Smart Structures and Systems, *Vol. 29, No. 1 (2022) 195-206*
https://doi.org/10.12989/sss.2022.29.1.195

An active learning method with difficulty learning mechanism for crack detection

Jiangpeng Shu [1a], Jun Li [1,2b], Jiawei Zhang [1], Weijian Zhao [1], Yuanfeng Duan [1] and Zhicheng Zhang*[1]

[1] *College of Civil Engineering and Architecture, Zhejiang University, 310058 Hangzhou, China*
[2] *Center for Balance Architecture, Zhejiang University, 310058 Hangzhou, China*

(*Received April 28, 2021, Revised July 28, 2021, Accepted August 3, 2021*)

Abstract. Crack detection is essential for inspection of existing structures and crack segmentation based on deep learning is a significant solution. However, datasets are usually one of the key issues. When building a new dataset for deep learning, laborious and time-consuming annotation of a large number of crack images is an obstacle. The aim of this study is to develop an approach that can automatically select a small portion of the most informative crack images from a large pool in order to annotate them, not to label all crack images. An active learning method with difficulty learning mechanism for crack segmentation tasks is proposed. Experiments are carried out on a crack image dataset of a steel box girder, which contains 500 images of 320×320 size for training, 100 for validation, and 190 for testing. In active learning experiments, the 500 images for training are acted as unlabeled image. The acquisition function in our method is compared with traditional acquisition functions, i.e., Query-By-Committee (QBC), Entropy, and Core-set. Further, comparisons are made on four common segmentation networks: U-Net, DeepLabV3, Feature Pyramid Network (FPN), and PSPNet. The results show that when training occurs with 200 (40%) of the most informative crack images that are selected by our method, the four segmentation networks can achieve 92%-95% of the obtained performance when training takes place with 500 (100%) crack images. The acquisition function in our method shows more accurate measurements of informativeness for unlabeled crack images compared to the four traditional acquisition functions at most active learning stages. Our method can select the most informative images for annotation from many unlabeled crack images automatically and accurately. Additionally, the dataset built after selecting 40% of all crack images can support crack segmentation networks that perform more than 92% when all the images are used.

Keywords: acquisition function; active learning; crack detection; probability attention module; semantic segmentation

1. Introduction

Surface cracking is a significant early indicator of structural damage. Visual inspection is often conducted to examine crack width, length, and distribution to ensure safety and durability of structures. Manual visual inspection is costly, time-consuming, subjective, and sometimes, dangerous as introduced by Kim *et al*. (2019) and Modarres *et al*. (2018). In these cases, automatic crack detection methods of image processing techniques (IPTs) based on computer vision are proposed and studied. Some of the well-known IPTs are the least squares method (Fujita *et al*. 2006), image binarization methods (Nguyen *et al*. 2014), percolation models (Oliveira and Correia 2009), and so on. However, most of the traditional image analysis methods focus on crack detection under non-complex conditions. For this reason, their applications are limited for engineering practice because of the great variation of image data that involve noise and complex background.

With the rapid development of machine learning and artificial intelligence, convolutional neural networks (CNNs) are applied to different visual understanding tasks, such as image recognition (He *et al*. 2016), object detection (Hadidi *et al*. 2014), and semantic image segmentation (Chen *et al*. 2017). Because of the success of CNN, more algorithms that are based on them have been proposed for structural damage inspection. As for the crack segmentation task, it is a binary classification problem of distinguishing "crack" and "non-crack" pixels. Semantic segmentation methods (Zhang *et al*. 2019) or pixel-level classification (Yang *et al*. 2018) are used to precisely delineate damage level, shape, and location. For pavement crack detection problems, an efficient model based on R-CNNs named CrackNet, was proposed (Pathirage *et al*. 2019, Zhang *et al*. 2019). Further, Xu *et al*. (2019) proposed a fatigue crack identification technique for long-span steel box girder bridges using deep CNN, in addition to a high-accuracy framework while using a restricted Boltzmann machine. Similarly, a U-Net based framework for concrete crack detection was developed (Liu *et al*. 2019), which can identify crack locations under various conditions with complex background with high efficiency and robustness. As mentioned above, many deep learning-based methods are extensively studied for applications to structural health monitoring (SHM), including the crack segmentation task (Fan *et al*. 2019, Ye *et al*. 2019).

*Corresponding author, Ph.D., Associate Professor,
E-mail: jszzc@zju.edu.cn
[a] Ph.D., E-mail: jpeshu@zju.edu.cn
[b] Ph.D. Student, E-mail: junli_zj@163.com

ISSN: 1738-1584 (Print), 1738-1991 (Online)

Based on deep learning, the core component of any application of segmentation methods is data (Bao *et al*. 2019); this also applies to the crack segmentation task. Currently, the number of available image databases of structural systems and other infrastructure components is very limited for SHM purposes; this leads to poor performance of the available trained models when new conditions arise, such as texture, joints, light, environment, and pollution (Azimi *et al*. 2020). Fortunately, the cost of portable devices and cameras has decreased; therefore, it is feasible to access and monitor parts of infrastructures through autonomous systems, such as unmanned aerial systems (UASs) (Kang and Cha 2018). Individuals can create datasets from their resources. For example, Yang *et al*. (2018) created a collection of 800 images of various cracks with 224 × 224 pixels to achieve semantically identification and pixel-wise segmentation.

However, building a new dataset is not easy. The manual annotation of crack images relied on experts, making it costly, time-consuming, and laborious, as mentioned before. Therefore, the high cost of annotation is an obstacle in increasing limited databases. In particular, the amount of unlabeled crack images is increasing dramatically with the increasing development of diversified image acquisition methods. To deal with this problem, active learning can be a promising solution by identifying the most informative crack images. Annotating these selected crack images to build a dataset can support sufficient supervision information and reduce the requirement of labeled crack images dramatically.

In the machine learning community, the attempt of active learning is to maximize a model's performance while annotating the fewest samples possible. (Ren *et al*. 2020) Active learning can be used as a method to reduce the cost of samples annotations, while retaining the powerful learning capabilities of deep learning. In the machine learning, both semi-supervised learning and active learning utilize the labeled and unlabeled samples on autonomous label engineering. Generally, semi-supervised learning does not require manual participation, and it can automatically label the unlabeled samples through a benchmark classifier with a certain classification accuracy. One of the characteristics of active learning, which is different from semi-supervised learning, is that the selected high-value samples need to be marked manually and accurately. Semi-supervised learning replaces manual labeling with automatic or semi-automatic labeling by computer. Although the labeling cost is effectively reduced, the labeling results depend on the classification accuracy of the benchmark classifier trained with some labeled samples, so the labeling results cannot be guaranteed to be completely correct. In contrast, active learning selects samples manually and does not introduce error class labels.

The first application of active learning in SHM is conducted by Wang *et al*. (2020). They proposed a sampling and training method based on active learning to treat class imbalances. In Wang's study, highly complex, yet informative images were successfully selected out and a very large amount of annotation work was saved. However, the crack detection in the study was based on the classification of crack images and non-crack images with a re-tuned AlexNet. Besides, the way of measuring the image information is an application of cross entropy, without a comparison of other active learning methods.

The active learning method can be divided into two groups: uncertainty-based (Wang *et al*. 2017, Cai and Wei 2020, Yoo and Kweon 2019) and representation-based (Sener and Savarese 2017, Sinha *et al*. 2019). The representation-based groups consider active learning as an approximation of the entire data distribution and query samples to increase data diversity, such as Core-set (Sener and Savarese 2017) and VAAL (Variational Adversarial Active Learning) (Sinha *et al*. 2019), which can be directly used in semantic segmentation. The main idea of representation-based methods attempts to use the core set to represent the distribution of the feature space of the entire original dataset, without reducing the diversity of dataset. The Core-set considers active learning as a problem to searching the best set from unlabeled data. The VAAL considers both the data distribution and the model uncertainty, to avoid the interference of abnormal data. The representation-based method is not particularly designed for segmentation tasks. However, it's still necessary to examine this kind of method for the crack segmentation task. In this study, Core-set, is used to yield crack images from the unlabeled dataset in the active learning experiments and its performance is discussed in detail in Section 4.

There are also some methods that are specifically designed for the semantic segmentation task. These methods can be divided into two groups: region-level (Mackowiak *et al*. 2018, Siddiqui *et al*. 2020) and image-level (Kuo *et al*. 2018, Wang *et al*. 2017, Yang *et al*. 2017). Region-level methods only sample the informative regions from images. By Mackowiak *et al*. (2018), MC dropout (Monte-Carlo dropout) uncertainty was combined with an effort estimation regressed from the annotation click patterns. Through hand-labeling a few, automatically selected, regions within an unlabeled image corpus, the annotation work was reduced. The regions of one image are the key to save label cost. For crack images, unlike the semantic datasets, there are no large regions for selection and cracks usually make up a small part of one image. Image-level methods calculate the complete image. Yang *et al*. (2017) proposed suggestive annotation (SA), trained a group of models on various labeled sets obtained with bootstrap sampling, and selected samples with the highest variance. It shows the potential of active learning in pixel-wise binary segmentation tasks. Kuo *et al*. (2018) adopted Query-By-Committee (QBC) strategy and proposed a cost-sensitive active learning method. The idea of QBC is to run multiple models on the same example and use their disagreement to estimate uncertainty. the cost-sensitive system is diverse and meaningful, maximizing the return on investment. It is more helpful when an image has a big ratio of the parts that need annotation, like bleeding in Kuo's biomedical segmentation work. The system is sensitive to the annotation cost and efficiency. For the crack segmentation task, since there is no such research on different active learning methods. The mentioned methods, including representation-based method Core-set,

uncertainty-based method Entropy, and QBC are applied in this study.

Inspired by the work of Xie *et al*. (2020), the semantic difficulty can be used to measure the informativeness and select samples at the image level for the crack segmentation tasks. The difficulty map can be learned with a self-attention mechanism. The typical crack geometrical characteristics, long and thin, get more attention and relatively receive a reasonable weighted information measurement. By Vaswani *et al*. (2017), the self-attention mechanism is first proposed for machine translation. Because of its intuition, versatility, and interpretability (Chaudhari *et al*. 2019), the self-attention mechanism is extensively studied for many tasks (Vaswani *et al*. 2017, Zhou *et al*. 2018). Many segmentation tasks adopt the attention module to capture long-range dependencies. Pan *et al*. (2020) not only considers the semantic interdependencies in spatial and channel dimensions, but also adaptively integrates local features into their global dependencies. Wang *et al*. (2018) designed two types of attention modules to exploit the dependencies between pixels and channel maps. The pixel-wise positional attention mechanism by Wang *et al*. (2018) that is used to aggregate similar pixels shows more details of targets and makes the object boundaries clearer, such as 'pole' and 'sidewalk'. Considering the geometrical similarity with crack, the position attention is adopted to improve segmentation results. Moreover, through the attention mechanism, with the obtained difficulty map, the calculation of one image uncertainty is improved. The hard part of crack segmentation, the edge of the crack, can get a higher weight in the information evaluation. Based on the selected crack images, the advantages of the proposed method which is especially designed for crack segmentation task are verified on four segmentation networks and are compared with several traditional active learning strategies.

2. Methods

Before introducing our method, the definition of an active learning problem is first given as follows. For an initial labeled dataset D^l, there are samples and the corresponding labels. For a much larger unlabeled dataset D^u, it contains unlabeled samples. Active learning aims to iteratively propose a subset D^s, which contains the most informative m samples selected from D^u (where m is a fixed constant, called 'Budget') (Xie *et al*. 2020).

In this study, an active learning method with a difficulty learning mechanism is introduced to select the most informative crack images. Considering the background diversities and shape disparity of cracks, a semantic difficulty difference exists among different semantic areas. To capture this difference, a two-branch network with difficulty learning mechanism is adopted in the method. One branch acts as a segmentation means, and the other one works for difficulty learning. The former segmentation branch adopts a common segmentation network. In our experiments, four segmentation models including U-Net (Ronneberger *et al*. 2015), PSPNet (Zhao *et al*. 2017), Feature Pyramid Network (FPN) (Lin *et al*. 2017), and DeepLabV3 (Chen *et al*. 2017), are used for evaluation. Through the segmentation branch, the probability map and prediction result are generated. After comparing the prediction results and ground truth, the wrong prediction is obtained. For the latter difficulty learning branch, the wrong prediction result is leveraged as the supervision. In the second branch, a pixel-wise attention module is adopted first to aggregate similar pixels, and learn the proportion of misclassified pixels as the difficulty score. In the last of the second branch, the semantic difficulty map is obtained. Based on the difficulty map and the uncertainty map from segmentation, an acquisition function is utilized to measure the informativeness of images in the unlabeled dataset D^u. These unlabeled images are ranked by their informative-

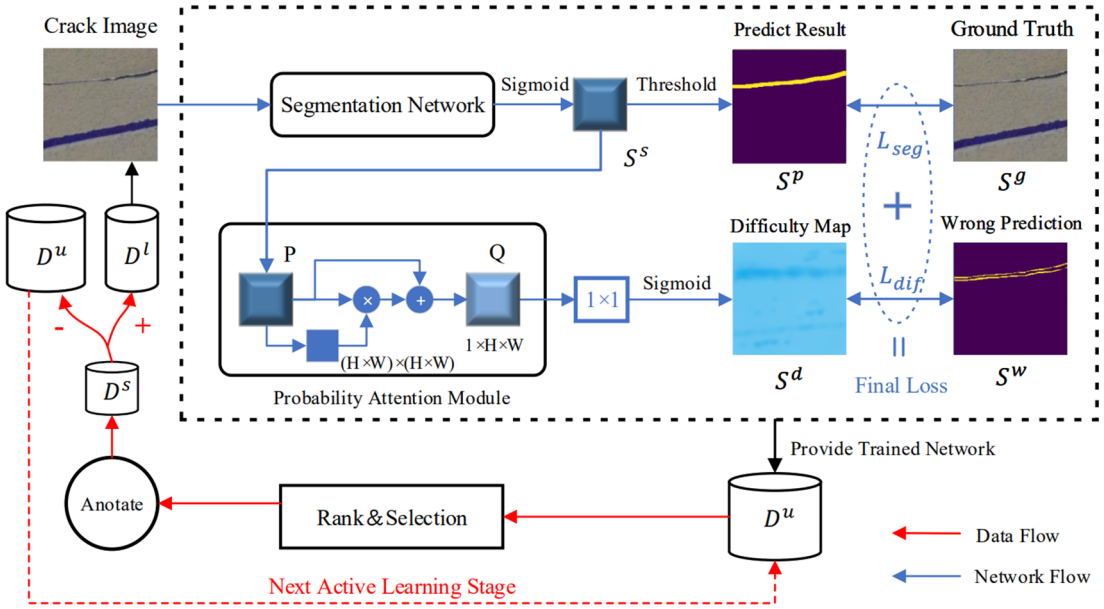


Fig. 1 Overview of the active learning method with difficulty learning mechanism

ness. The most informative m samples of them are selected. After annotation by experts, the selected m samples form a subset D^s. This subset is added to the labeled dataset D^l, and subtracted from D^u. The final D^u is then ready for the next active learning stage. The overview of the active learning method with the learning mechanism is shown in Fig. 1.

Subsequently, a more detailed active learning framework is introduced. Then, the probability attention module and loss functions are presented. Finally, the acquisition function combined with learned difficulty maps is given.

2.1 Active learning framework with difficulty learning mechanism

A two-branch network is built to complete a multi-task manner, and generates crack segmentation predictions and difficulty maps. The structure of the network is displayed in Fig. 1. The first branch is a common segmentation network. As illustrated in Fig. 1, the output of the sigmoid layer is assumed as S^s, and the prediction result S^p is generated at the threshold of *0.5*. By comparing S^p and the ground truth S^g, the wrong prediction result S^w that will be used at the second branch to learn semantic difficulty, is computed using the following equation

$$S_i^w = \begin{cases} 1 & if\ S_i^P \neq S_i^g \\ 0 & if\ S_i^P = S_i^g \end{cases} \tag{1}$$

where S_i^w is the i^{th} pixel value of the wrong prediction result, and S_i^P and S_i^g represent the i^{th} pixel value of prediction result and ground truth, respectively.

The second branch comprises two parts. The first part is a probability attention module. The sigmoid output of the first branch S^s is directly used as the input of the second branch. For the crack segmentation task, S^s is a one-channel probability map. Thus, the input of the second branch, named P, is the shape of C × H × W (Chanel × Height × Width). There is an advantage by directly using probability maps, because pixels with similar difficulty are likely to have a similar probability vector. If they are combined with a pixel-wise attention module, similar pixels are easily aggregated and similar difficulty scores are attached to them. The detailed attention module will be further illustrated in Section 2.2. The second part of the second branch is a simple 1 × 1 convolution layer to process the output of the probability attention module for binary classification.

In conclusion, the whole semantic difficulty learning process can be divided into two steps. First, a learned segmentation network is trained with the wrong prediction results S^w. Thereafter, every pixel will learn a semantic difficulty value. Finally, the semantic difficulty map S^d is obtained.

2.2 Probability attention module

The probability attention module in our method is explained in detail in this section. According to Fu *et al.* (2019), pixels with similar probability will be aggregated through this module. The input of the second branch, i.e., the probability map P, is the shape of $C \times H \times W$. P is first reshaped into $C \times K$, in which K is calculated by $H \times W$. Then, the attention matrix A, which is the shape of K × K, is generated with $P^T P$ and a sigmoid operation as follows

$$A_{ji} = \frac{exp\ (P_i^T \cdot P_j)}{\sum_{i=1}^{K} exp\ (P_i^T \cdot P_j)} \tag{2}$$

$$Q_j = \gamma \sum_{i=1}^{K} \left(A_{ji} P_i\right) + P_j \tag{3}$$

where A_{ji} is the i^{th} pixel's impact on the j^{th} pixel, P_j is the probability vector of the j^{th} pixel in the probability map P, Q_j is the result after attention γ is a learnable weight factor. The final probability map after the attention module (Q) is obtained.

Moreover, the probability attention module (Fu *et al.* 2019) is designed to capture the position relationship of pixels. The boundaries of objects, like 'pole', are paid more attention and predicted more precisely. The crack has a similar geometrical characteristic with 'pole'. The edge of cracks also is a key point and hard part for pixel-wise segmentation. The promotion of probability in the method is verified through an ablation study by removing this module. The detailed discussion is shown in Section 4.2.

2.3 Loss functions

2.3.1 Loss of semantic segmentation

In the crack segmentation task, one crack image usually contains only 3% pixels that belong to the crack, which is particularly unbalanced. According to Milletari *et al.* (2016), Dice loss contributes to datasets that contain unbalanced samples. Moreover, it can improve the main certification standard (IoU), and it is a proper choice for our crack segmentation task. It is defined as

$$L_{seg}(S^p, S^g) = \frac{2\sum_{i=1}^{N} S_i^p S_i^g}{\sum_{i=1}^{N} (S_i^p)^2 + \sum_{i=1}^{N} (S_i^g)^2} \tag{4}$$

where S^p, S^g are the segmentation output and ground truth, respectively; S_i^p, S_i^g are the values for the i^{th} pixel, and N is the total pixel number.

2.3.2 Loss of semantic difficulty

For the semantic difficulty branch (the second branch), an inverted weighted binary cross-entropy loss is used, while considering the imbalance between the right and wrong areas of error mask. It is defined as

$$L_{dif}(S^d, S^w) = -\frac{1}{N}\sum_{i=1}^{N} \lambda_1 S_i^w log\left(S_i^d\right) + \lambda_2 (1 - S_i^w) log\left(1 - S_i^d\right) \tag{5}$$

$$\lambda_1 = \frac{\sum_i^N f(S_i^w = 0)}{N}, \qquad \lambda_2 = 1 - \lambda_1 \tag{6}$$

where S^d, S^w are the difficulty map and wrong prediction result map; S_i^d, S_i^w are the ith pixel value, N is the total pixel number, $f(\cdot)$ is the indicator function, and λ_1 and λ_2 are dynamic weight factors.

2.3.3 Final loss

The network is trained with the final loss, considering both the segmentation and difficulty loss. The final training loss is a simple combination of Eqs. (5) and (6). It is computed as

$$L = L_{seg}(S^p, S^g) + L_{dif}(S^d, S^w) \tag{7}$$

2.4 Acquisition functions

In active learning, samples from the unlabeled dataset D^u will be scored and ranked using different acquisition functions. If the traditional acquisition function is combined with the difficulty maps yielded in network, it can be improved. Normally, the traditional functions are considering pixels in an average way and are not concerned that different pixels have different difficulty to learn. For crack segmentation, the difficulty map is utilized to add weights to different pixels. In this way, pixels that are considered as hard to learn will be attached to higher weights, and pixels which are easy to learn will get lower weights. Assuming that M^t is the uncertainty map generated with traditional acquisition function and S^d is the difficulty map, the acquisition function for one image is given as follows

$$F = \frac{1}{N}\sum_{i=1}^{N} M_i^t S_i^d \tag{8}$$

where M_i^t, S_i^d are the traditional uncertainty score and difficulty value of the ith pixel, respectively; N is the total pixel number, and F is the new uncertainty score of one image to select samples with the most informativeness.

2.5 Processing flow

The steps of the processing flow are given in the following list:

(1) Acquire an unlabeled database.
(2) Select randomly an initial 10% of the samples from the unlabeled database. Annotate these samples to form D^l, and use the rest of the unlabeled samples to form D^u.
(3) Take D^l, D^u as the input, set the constant budget number m, the active learning query times N, the initialized network parameter Θ, and the iterations T.
(4) Train the two-branch network on D^l for T iterations. In one iteration, with the samples from D^l, the segmentation output S^s and prediction result S^p are obtained by the first branch. Then, according to Eq. (1), S^w is computed. The difficulty prediction S^d is proposed by the second branch. According to Eq. (7), the final loss L is calculated. In the last, the network parameter Θ is updated using gradient descent.
(5) Rank unlabeled samples from D^u using the trained network from step 4, according to Eq. (8).
(6) Select the top m samples from D^u. Annotate them to form the subset D^s.
(7) Add D^s to D^l, and subtract D^s from D^u.
(8) Back to step 3, repeat steps 3 to 7 for N times.
(9) Get the final D^l, D^u, and trained network parameter Θ.

3. Experiments

In this section, the evaluation dataset is first introduced. Then, some additional remarks are made to the implementation details. The structures of U-Net, DeepLabV3, FPN, and PSPNet are simply presented. The five acquisition functions that are used for comparison, (Random, Entropy, QBC, Core-set, and ours) are introduced. The results of four networks trained with full training images are discussed. After the basic performance of the four networks on crack segmentation is obtained, the active learning experimental results at different stages are presented and discussed. Lastly, the experimental results for removing the probability attention module are discussed.

3.1 Evaluation dataset

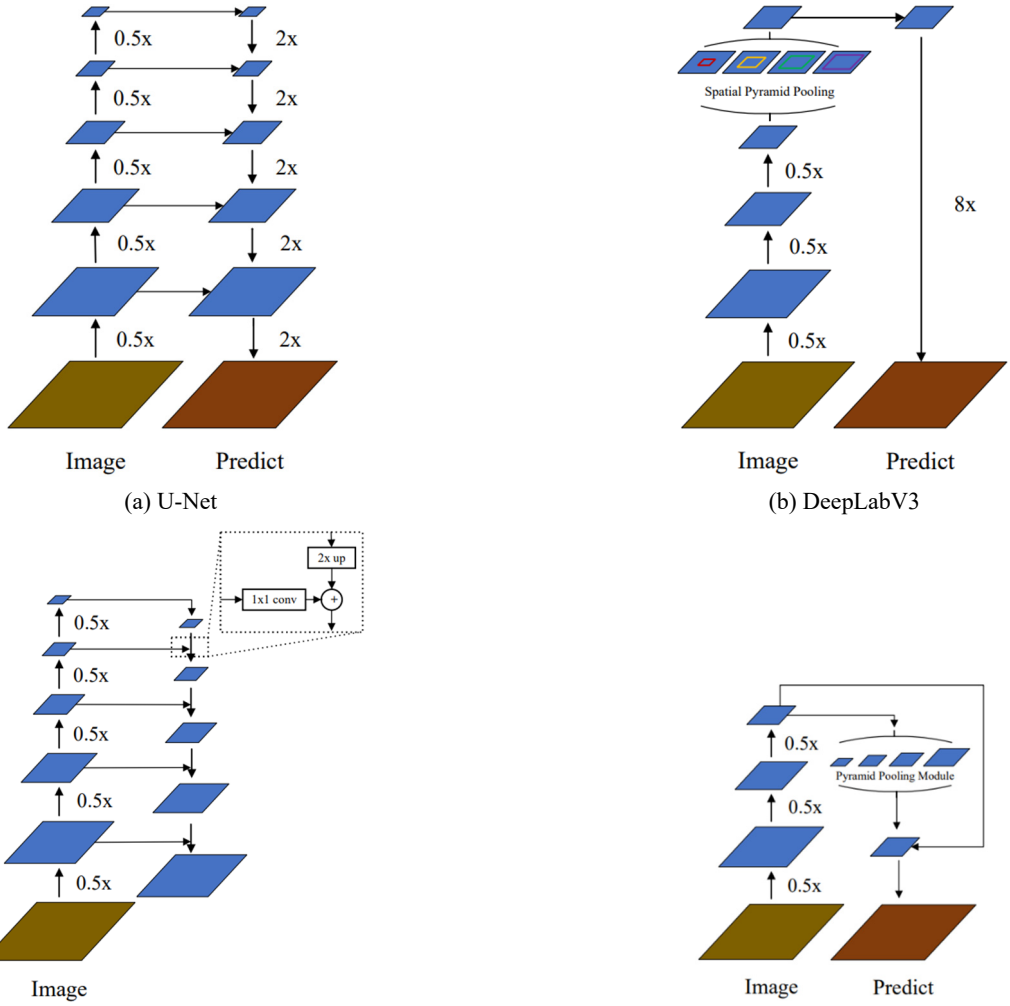
The crack images are provided by the International Project Competition for SHM (IPC-SHM 2020) ANCRiSST (Bao *et al*. 2021). There are 80 crack images with a resolution of 4396×3928. These images contain the crack information inside a steel box girder. They are labeled by experienced experts; the 80 images have corresponding labels. Because of the limitation of computer resources, it is difficult to deal with these high-resolution images directly. For convenience, the 80 images are cropped into much smaller images of 320×320. Finally, 790 small samples containing cracks are obtained. Some samples are shown in Fig. 4. Of these 790 samples, we randomly select 100 samples as the validation set, and 190 samples as the test set. The remaining 500 samples are treated as the training set. The detailes are listed in Table 1. Following the active learning process, 10% of the data are randomly selected from the training set as the initial annotated dataset, then iteratively query 5% of new data from the remaining training set, which serves as the unlabeled dataset. The percentage of every iteration and the initial ratio are constant in the active learning experiment, to make sure it not fine re-tuned, which makes the method more applicable.

3.2 Implementation details

When constructing the segmentation network in the first

Table 1 Details of evaluation dataset

Classes	Train	Valid	Test	Initial labeled	Budget	Image size
1	500	100	190	50	25	320 × 320



(a) U-Net (b) DeepLabV3

(c) FPN (d) PSPNet

Fig. 2 The encoding-decoding structures of evaluated segmentation networks

branch, a uniform structure of encoding-decoding is adopted. The ResNeXt-50 (Xie *et al*. 2017) is used as the encoding structure to extract features. The decoding part is different for the four networks in the experiment. In training, mini-batch SGD (Zhang *et al*. 2020) with momentum of *0.9* and weight decay of $5e^{-4}$ is used. The batch size is *10*. For all networks, acquisition functions and the upper bound method with the full training data, the networks are trained for *100* epochs. Similarly to (Chen *et al*. 2018), the "poly" learning rate strategy are applied and the initial learning rate is *0.0001* and multiplied by $\left(1-\frac{iter}{total\_iter}\right)^{0.9}$. To accelerate the calculation of the probability attention module, the input of the difficulty branch is resized to 80 × 80, rather than the original 320 × 320. The tradition acquisition function that is combined with the difficulty map is Entropy.

3.3 Evaluated segmentation models

All the experiments are relying on U-Net, DeeplabV3, FPN, and PSPNet. The characteristics and comparisons of the four networks are given as follows:

- U-Net is a traditional and legendary fully CNN for semantic image segmentation. The encoding and decoding parts are connected with skip connections. The encoding part extracts features of different spatial resolution. These features are then used by the decoder to generate a segmentation mask. Typically, U-Net uses concatenation for fusing decoder blocks with skip connections (see Fig. 2(a)).
- DeepLabV3 is a widely used segmentation network. It adopts modules that employ atrous convolution in cascade or in parallel to capture multi-scale context by using multiple atrous rates. Furthermore, the atrous spatial pyramid pooling (ASPP) module is augmented in DeepLabV3. The ASPP module probes convolutional features at multiple scales, with image-level features encoding global context and further boosting performance. So, DeepLabV3 attains comparable performance (see Fig. 2(b)).
- Feature Pyramid Network (FPN) was the winning entry in the COCO stuff 2017 competition. FPN exploits the inherent multi-scale pyramidal hierarchy of deep convolutional networks to construct feature pyramids with extra marginal cost. The lateral

connections are developed for building high-level semantic feature maps at all scales. Compared with PSPnet, FPN focuses on all features extracted by the decoder and shows significant improvement (see Fig. 2(c)).

- Pyramid Scene Parsing Net (PSPNet) adopts a spatial pyramid pooling structure and demonstrates outstanding performance on several semantic segmentation benchmarks. PSPNet consists of an encoder and a spatial pyramid (decoder). The spatial pyramid builds on the top of the encoder, and only uses features of low spatial resolution. Since PSPNet ignores features of high spatial resolution, it is not suitable for detecting small objects and producing precise pixel-wise masks (see Fig. 2(d)).

Additionally, as mentioned in section 4.2, the decoder of the four segmentation networks is the same, ResNeXt-50 (Xie *et al*. 2017). ResNeXt-50 and ResNet-50 (He *et al*. 2016) have similar numbers of parameters, but ResNeXt-50 shows better performance and only needs a few hyper-parameters to set. Thus, ResNeXt-50 is chosen as the encoder in the experiments. The pretrained encoder weight is "ImageNet." In Fig. 2, the encoding-decoding structures of the four segmentation models are illustrated.

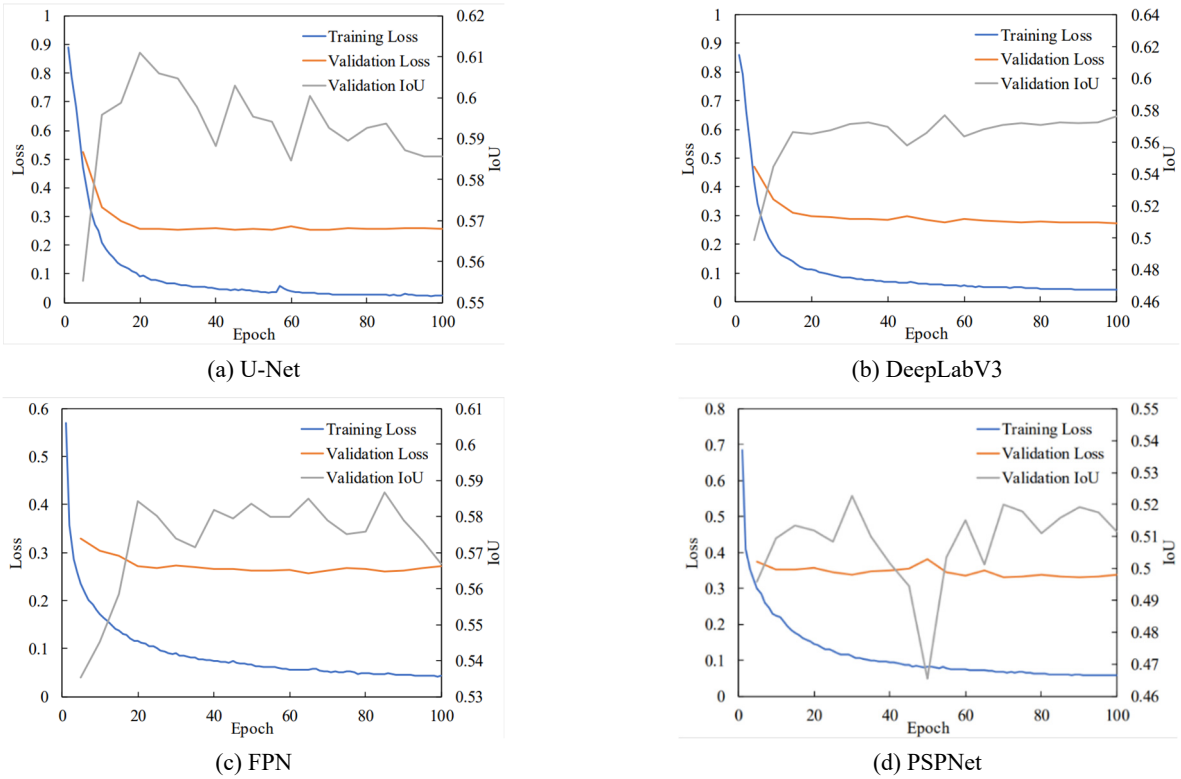
3.4 Evaluated acquisition functions

Five acquisition functions are used for comparison. Random is a simple baseline method; Entropy and QBC are two uncertainty-based methods. Core-set is a representation-based method. The acquisition function we suggest is combined with difficulty maps. All of them are presented in the following list:

- Random: each sample in D^u is queried with uniform probability, which is selected randomly.
- Entropy (Uncertainty): samples are queried with the maximum mean entropy of all pixels. This method is verified by Yoo and Kweon (2019), and is quite competitive for image classification and segmentation tasks.
- QBC (Uncertainty): methods that are designed for semantic segmentation (Mackowiak *et al*. 2018, Yang *et al*. 2017), use a group of models to measure uncertainty. The efficient MC dropout is used to represent these methods and report the best performance out of both the maximum-entropy and variation-ratio acquisition functions.
- Core-set (Representation): samples that can cover best the entire data distribution are queried. A global average pooling operation is applied on the encoder output features of ResNeXt-50, and a feature vector for each sample is obtained. Then according to Sener and Savarese (2017), k-center algorithm is used.
- Ours: the acquisition function in our method is the combination of a traditional function and difficulty maps. In the experiments, the Entropy is chosen as the traditional function. The final acquisition function is calculated by Eq. (8).

4. Results and discussions

The crack pixels only occupy 2%-3% of one image.



(a) U-Net (b) DeepLabV3

(c) FPN (d) PSPNet

Fig. 3 Training loss, validation loss, and validation IoU of U-Net, DeepLabV3, FPN, and PSPNet

Table 2 Test results of networks trained with full training data

Network	U-Net	DeepLabV3	FPN	PSPNet
IoU	0.625	0.613	0.607	0.545
FWIoU	0.901	0.893	0.886	0.869
PA	0.922	0.909	0.895	0.877
MPA	0.931	0.917	0.923	0.885

Pixel accuracy is not very suitable for the crack segmentation task. Since crack segmentation is a pixel-wise binary classification problem, the evaluation index intersection over union (IoU) is adopted. Every segmentation network and method are run for 5 times, and the average IoUs are reported. In the following subsections, the upper bound results with full training dataset on four segmentation networks are first discussed; thereafter, the active learning results of different methods are presented.

4.1 Upper bound results with full training data

Fig. 3 shows the training history of four segmentation models on the full training data (500 crack images). The trend of training loss and validation loss indicate that 100 epochs are enough for network to learn crack features. The validation IoU in the training is also shown in the Fig. 3. Table 2 shows the test results of four networks, including IoU, FWIoU (Frequency Weighted Intersection over Union), PA (Pixel Accuracy), MPA (Mean Pixel Accuracy). U-Net achieves the best IoU. For the other three networks with similar pyramid structures, DeepLabV3 shows a bit of advantage of FPN; FPN is 11.4% better than PSPNet. As an early proposed pyramid structure, PSPNet is known as not being good for small objects. This phenomenon is also observed on crack segmentation. FPN makes some adjustments in the decoding structure. The strengthened connection of lower and upper features helps remarkably to yield precise segmentation prediction. DeepLabV3 is famous for the spatial pyramid structure. It forms a strong sense of both small and large objects, achieving better results on COCO and other open datasets compared to FPN. FPN also shows better performance than PSPNet on open datasets in segmentation competitions. It indicates that networks, which have a strong sense of upper and lower features, can obtain good scores on open datasets/ competitions; they can also reach a good IoU remark on the crack segmentation task.

Meanwhile, by connecting every feature level tightly with skip connections, U-Net becomes the best network structure among the four evaluated segmentation networks.

The visual presentation of the segmentation results being predicted by different networks is shown in Fig. 4. As shown in Fig. 4, crack images of the first three rows have simple backgrounds. Prediction results of the four networks are fairly accurate, with only those of PSPNet showing some noise points. For the fourth row of Fig. 4, a crack in the image is covered by a blue marking pen. Being disturbed by human interference, this part of the crack is not successfully predicted by all four networks. In the future, it is expected to add more of this type of crack images for training to improve predictions. For the last row of Fig. 4, the crack image has a noticeable change in brightness. The networks U-Net and DeepLabV3 handle this with a good

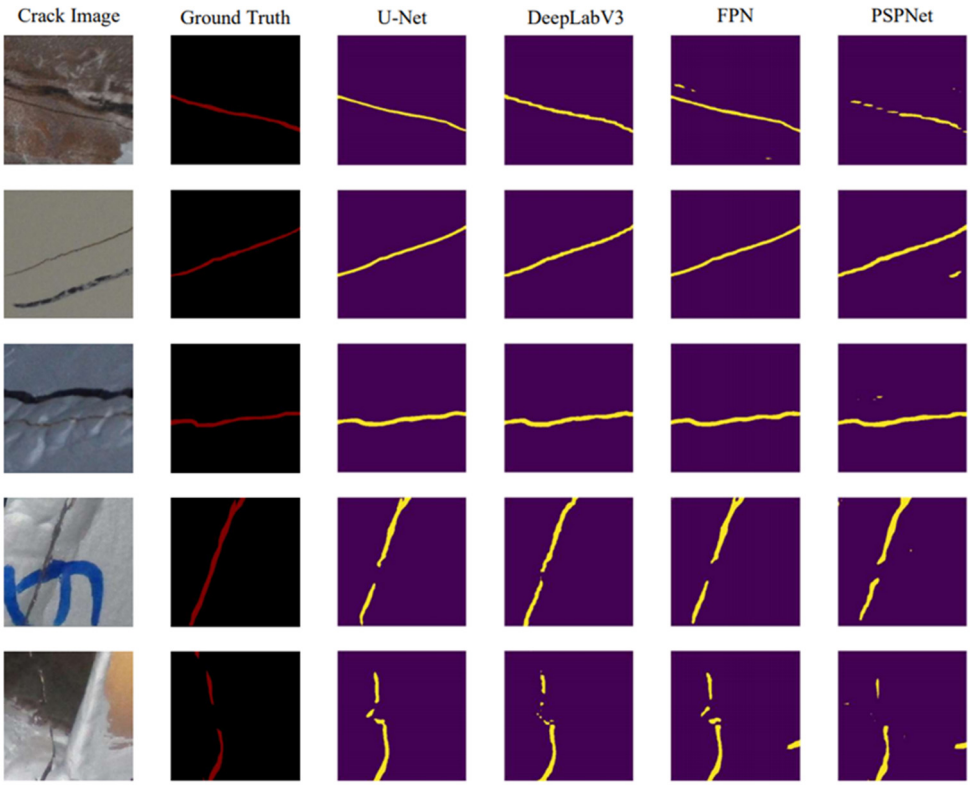
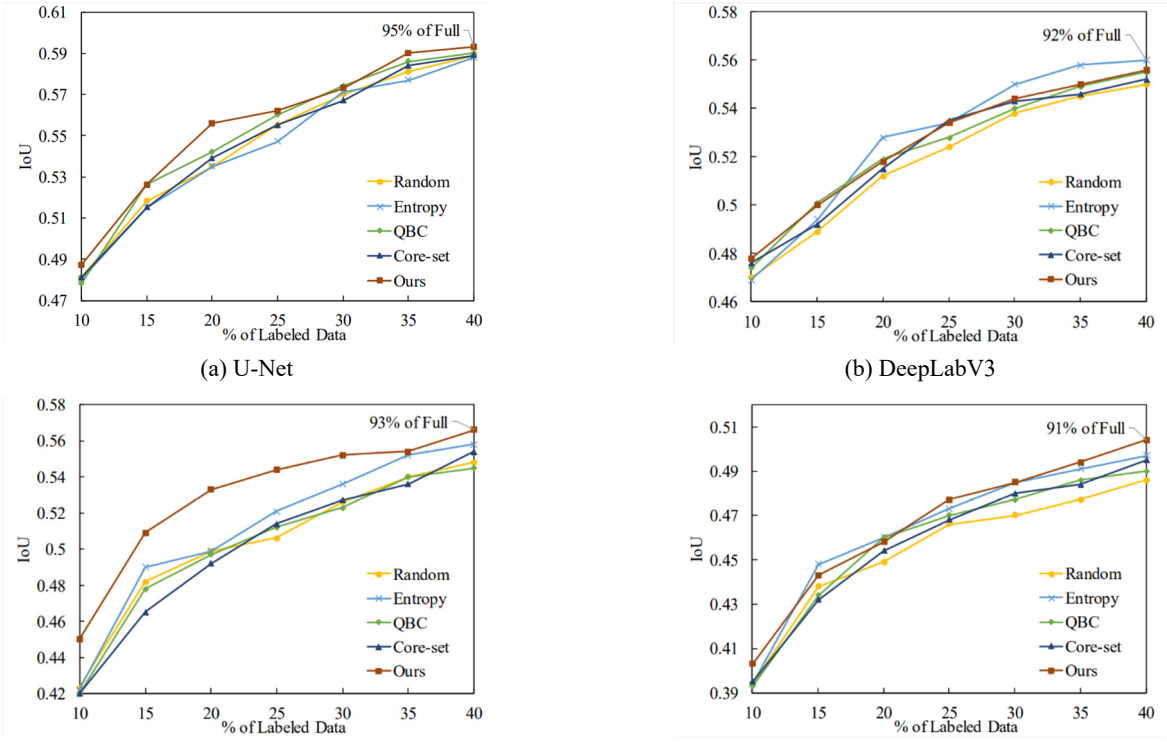


Fig. 4 Representative crack images and predictions of U-Net, DeepLabV3, FPN, and PSPNet



(a) U-Net (b) DeepLabV3

(c) FPN (d) PSPNet

Fig. 5 Active learning stages on U-Net, DeepLabV3, FPN, and PSPNet

performance. For FPN and PSPNet, some pixels are predicted incorrectly as cracks, and a crack in the dark part is not predicted precisely. In conclusion, the visual presentation in Fig. 4 is corresponding to the test results shown in Table 2. For certain situations involving human interference and complex background, it is possible to make improvements.

4.2 Comparison results on active learning

The test IoU at each active learning stage: 10%, 15%, 20%, 25%, 30%, 35%, 40% of the full training set, are recorded as evaluation metric. Depending on four segmentation networks, every method is run 5 times and the average IoUs are reported, as mentioned previously.

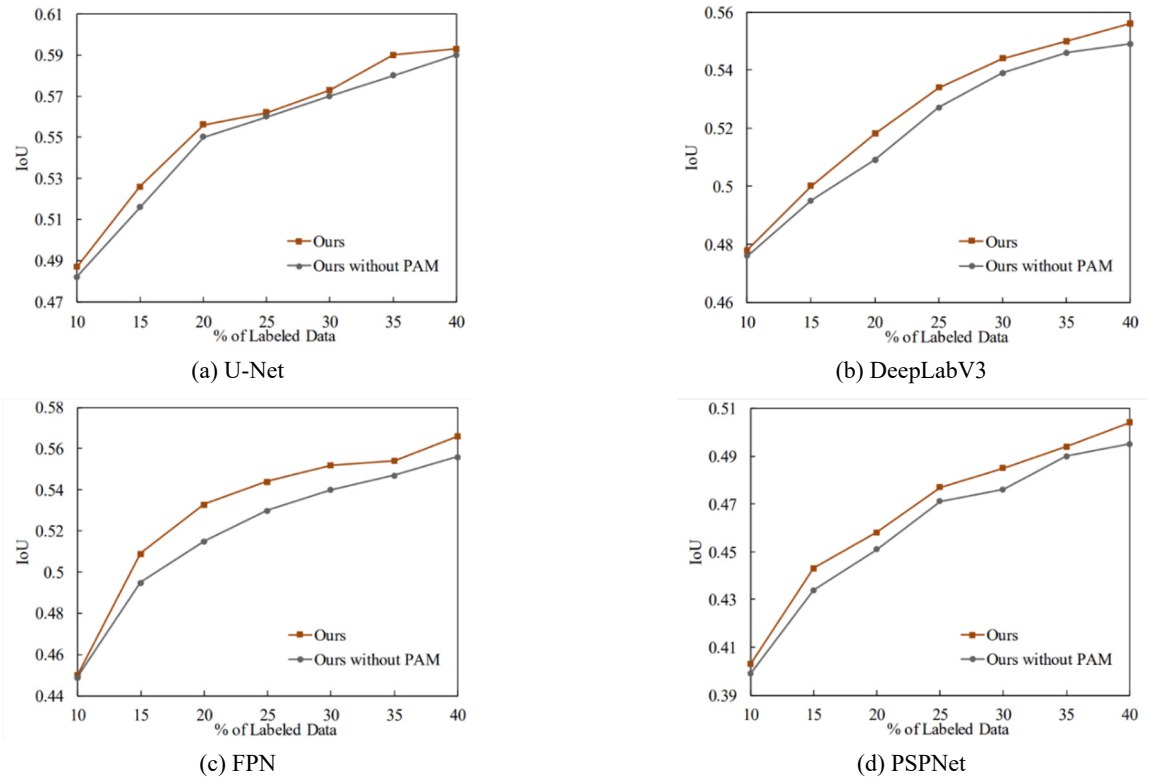
The comparison results on active learning of the four networks are shown in Fig. 5. The acquisition function that is combined with difficulty maps (Ours) outperforms the baseline function (Random) at all active learning stages on the four networks. For earlier stages like 15% and 20%, the performances of the four networks increase very quickly. For later stages, the rising rate slows down. At the final stage of active learning, when using 40% of the selected training samples, an average performance of the upper bound training results of 93% is obtained. However, for different networks, the superiority of the acquisition function combined with difficulty maps differs.

U-Net shows outstanding performance on full training data. Providing much less training data does not affect much its segmentation accuracy. At the initial stage, for 5% of full training samples, U-Net still obtains 80% of the upper method, which is the highest percentage among the four segmentation models. Through different functions, training budgets containing different informative level samples are generated. U-Net learns well from these new samples without being affected much by the level of informativeness. As shown in Fig. 5, at different active learning stages, various functions get very similar results.

For DeepLabV3 (see in Fig. 5), the acquisition function combined with difficulty maps is beyond the baseline function from the beginning to the end; but, it is not in the lead at every active learning stage. At the early active learning stage of 15%, the function combined with the difficulty map and traditional method Entropy show similar performance. At the later active learning stages of 30% and 35%, the Entropy has a higher growth rate and maintains the leading superiority at the end of active learning. For DeepLabV3, the Entropy method is competitive, but the acquisition function combined with difficulty map does not fall behind too much and it is still suggested.

For FPN (see in Fig. 5), the most representative improvement is observed. At early stages of 15% and 20%, it performs 10.6% and 5.6% better than the baseline method. At the later stages of 35% and 40%, the function we suggest has a more stable and smooth growth curve. In the last stage of 40%, it maintains the best performance. When considering the other functions, the Entropy method also shows excellent performance. At the early active learning stages, it does not catch up with the function combined with the difficulty map. At the final stage with more training samples, the segmentation ability of FPN makes up for the shortage of informativeness.

PSPNet shows the weakest performance on the crack segmentation task on full training data. In Fig. 5, at the



(a) U-Net (b) DeepLabV3

(c) FPN (d) PSPNet

Fig. 6 Comparison experiments of removing probability attention module (PAM) on U-Net, DeepLabV3, FPN, and PSPNet

initial active learning stage of 10%, all methods that depend on PSPNet get the lowest IoU among the four segmentation networks; this corresponds to the upper bound result trend. Due to the weakness of PSPNet on crack segmentation at the later active learning stage, the improvement of different methods is not significant. The similar results of all acquisition functions indicate that the amount of training samples is more significant for raising IoU than the informativeness level of training samples.

For different acquisition functions, their performance is influenced by the segmentation ability of different networks. Except for the function that is combined with the difficulty map (Ours), different functions have a changing advantage. Only the function used in the framework maintains superior performance at most active learning stages and gets the highest remark at the last stage. Thus, on active learning for crack segmentation, the acquisition function that is combined with difficulty map is recommended.

To further understand the effect of probability attention module, an ablation study is conducted by removing this module. Four segmentation networks directly learn the semantic difficult map without the attention among pixels. Without the long-range dependence, pixels of the same semantic can learn quite different scores because the learned score of each pixel tends to be more sensitive to the original uncertainty value. If it is combined with the probability attention module, a smoother difficulty map is learned; this is better for the annotators since the aggregated semantic areas are close to the labeling units in the real scenario. The comparison results between the ablation models and the original ones are shown in Fig. 6. For all the segmentation networks, the acquisition function with the probability attention module can achieve better performance at each active learning stage. Without the module, it becomes hard to find samples with more balanced semantic difficulty. Through the ablation study, the benefit of the probability attention module is verified. For different segmentation networks, the probability attention module is suggested to be used in order to improve active learning.

5. Conclusions

In this study, a difficulty-learning active learning method to select the most informative crack images is proposed, aiming at promoting crack segmentation dataset construction. An acquisition function that is combined with difficulty and traditional uncertainty maps is utilized to measure the informativeness of crack images in our method. A steel box girder crack image dataset containing 500 images of 320 × 320 pixels for training, 100 for validation, and 190 for testing is used for our experiments. Four common segmentation networks, including U-Net, DeepLabV3, FPN, and PSPNet, are applied to segment cracks. A comparison study between our acquisition function and traditional acquisition functions, including Random, QBC, Entropy, and Core-set, is conducted on the four networks. The role of probability attention module (PAM) that is used in our method is evaluated by experiments through removing this module.

- The most informative crack images for annotation can be selected automatically and accurately using our method. Using these selected crack images to

build a dataset for training can provide sufficient information for the crack segmentation network.

- At the final stage of active learning, when using 200 (40%) samples selected by our method the four evaluated segmentation networks can obtain 92%–95% of the performance when using 500 (100%) crack images for training.
- Additionally, the comparison study indicates that the proposed acquisition function has more accurate measurements for unlabeled crack images. The evaluation index IoU of our acquisition function is the highest at most active learning stages.
- The performance of acquisition functions is affected by segmentation ability of networks on crack. The relatively weak network on crack segmentation is suggested to increase training sample amounts. More informative samples are recommended to a relatively strong network.
- The results of U-Net, DeepLabV3, FPN, and PSPNet being trained with 500 (100%) crack images demonstrate that networks with stronger connections between upper and lower features in the decoder can yield more precise crack predictions.
- The experiments of removing the PAM indicate that the module is recommended for the active learning method on the four segmentation networks at all active learning stages.

Acknowledgments

The authors would like to thank the organizations of the International Project Competition for SHM (IPC-SHM 2020) ANCRiSST, Harbin Institute of Technology (China), and University of Illinois at Urbana-Champaign (USA) for their generously providing the invaluable data from actual structures. The authors would also like to gratefully acknowledge the support from the Nation Natural Science Foundation of China (52108179, U1709216), the China Postdoctoral Science Foundation (2021M692835, 2021M702866), and the National Key R&D Program of China (2018YFE0125400), which made the research possible.

References

Azimi, M., Eslamlou, A.D. and Pekcan, G. (2020), "Data-driven structural health monitoring and damage detection through deep learning: state-of-the-art review", *Sensors*, **20**(10), 2778. https://doi.org/10.3390/s20102778

Bao, Y., Chen, Z., Wei, S., Xu, Y., Tang, Z. and Li, H. (2019), "The state of the art of data science and engineering in structural health monitoring", *Engineering*, **5**(2), 234-242. https://doi.org/10.1016/j.eng.2018.11.027

Bao, Y., Li, J., Nagayama, T., Xu, Y., Spencer Jr., B.F. and Li, H. (2021), "The 1st International Project Competition for Structural Health Monitoring (IPC-SHM, 2020): A summary and benchmark problem", *Struct. Health Monitor.*, **20**(4), 2229-2239. https://doi.org/10.1016/j.eng.2018.11.027

Cai, W. and Wei, Z. (2020), "Remote sensing image classification based on a cross-attention mechanism and graph convolution", *IEEE Geosci. Remote Sens. Lett.* https://doi.org/10.1109/LGRS.2020.3026587

Chaudhari, S., Polatkan, G., Ramanath, R. and Mithal, V. (2019), "An attentive survey of attention models", arXiv preprint arXiv: 1904.02874. https://arxiv.org/abs/1904.02874

Chen, L.-C., Papandreou, G., Schroff, F. and Adam, H. (2017), "Rethinking atrous convolution for semantic image segmentation", arXiv preprint arXiv: 1706.05587. https://arxiv.org/abs/1706.05587

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018), "Encoder-decoder with atrous separable convolution for semantic image segmentation", *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, Germany, September.

Fan, G., Li, J. and Hao, H. (2019), "Lost data recovery for structural health monitoring based on convolutional neural networks", *Struct. Control Health Monitor.*, **26**(10), 1-21. https://doi.org/10.1002/stc.2433

Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z. and Lu, H. (2019), "Dual attention network for scene segmentation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June.

Fujita, Y., Mitani, Y. and Hamamoto, Y. (2006), "A method for crack detection on a concrete structure", *Proceedings of International Conference on Pattern Recognition*, Hong Kong, China, August.

Hadidi, N.N., Cullen, K.R., Hall, L.M.J., Lindquist, R., Buckwalter, K.C. and Mathews, E. (2014), "Functional magnetic resonance imaging as experienced by stroke survivors", *Res. Gerontol. Nurs.*, **7**(5), 200-205. https://doi.org/10.3928/19404921-20140820-01

He, K., Zhang, X., Ren, S. and Sun, J. (2016), "Deep residual learning for image recognition", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, June.

Kang, D. and Cha, Y.J. (2018), "Autonomous uavs for structural health monitoring using deep learning and an ultrasonic beacon system with geo-tagging", *Comput.-Aided Civil Infrastr. Eng.*, **33**(10), 885-902. https://doi.org/10.1111/mice.12375

Kim, H., Ahn, E., Shin, M. and Sim, S.H. (2019), "Crack and noncrack classification from concrete surface images using machine learning", *Struct. Health Monitor.*, **18**(3), 725-738. https://doi.org/10.1177/1475921718768747

Kuo, W., Häne, C., Yuh, E., Mukherjee, P. and Malik, J. (2018), "Cost-sensitive active learning for intracranial hemorrhage detection", *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, Granada, Spain, September.

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B. and Belongie, S. (2017), "Feature pyramid networks for object detection", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July.

Liu, Z., Cao, Y., Wang, Y. and Wang, W. (2019), "Computer vision-based concrete crack detection using U-net fully convolutional networks", *Automat. Constr.*, **104**, 129-139. https://doi.org/10.1016/j.autcon.2019.04.005

Mackowiak, R., Lenz, P., Ghori, O., Diego, F., Lange, O. and Rother, C. (2018), "Cereals-cost-effective region-based active learning for semantic segmentation", arXiv preprint arXiv: 1810.09726. https://arxiv.org/abs/1810.09726

Milletari, F., Navab, N. and Ahmadi, S.A. (2016), "V-Net: Fully convolutional neural networks for volumetric medical image segmentation", *Proceedings of International Conference on 3D Vision*, Stanford, CA, USA, October.

Modarres, C., Astorga, N., Droguett, E.L. and Meruane, V. (2018), "Convolutional neural networks for automated damage

recognition and damage type identification", *Struct. Control Health Monitor.*, **25**(10), 1-17. https://doi.org/10.1002/stc.2230

Nguyen, H.N., Kam, T.Y. and Cheng, P.Y. (2014), "An automatic approach for accurate edge detection of concrete crack utilizing 2d geometric features of crack", *J. Signal Process. Syst.*, **77**(3), 221-240. https://doi.org/10.1007/s11265-013-0813-8

Oliveira, H. and Correia, P.L. (2009), "Automatic road crack segmentation using entropy and image dynamic thresholding", *Proceedings of European Signal Processing Conference*, Glasgow, Scotland, UK, August.

Pan, Y., Zhang, G. and Zhang, L. (2020), "A spatial-channel hierarchical deep learning network for pixel-level automated crack detection", *Automat. Constr.*, **119**, 103357. https://doi.org/10.1016/j.autcon.2020.103357

Pathirage, C.S.N., Li, J., Li, L., Hao, H., Liu, W. and Wang, R. (2019), "Development and application of a deep learning–based sparse autoencoder framework for structural damage identification", *Struct. Health Monitor.*, **18**(1), 103-122. https://doi.org/10.1177/1475921718800363

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Chen, X. and Wang, X. (2020), "A survey of deep active learning", arXiv preprint arXiv: 2009.00236. https://arxiv.org/abs/2009.00236

Ronneberger, O., Fischer, P. and Brox, T. (2015), "U-Net: convolutional networks for biomedical image segmentation", *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, Munich, Germany, October.

Sener, O. and Savarese, S. (2017), "Active learning for convolutional neural networks: A core-set approach", arXiv preprint arXiv: 1708.00489. https://arxiv.org/abs/1708.00489

Siddiqui, Y., Valentin, J. and Nießner, M. (2020), "Viewal: Active learning with viewpoint entropy for semantic segmentation", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, USA, June.

Sinha, S., Ebrahimi, S. and Darrell, T. (2019), "Variational adversarial active learning", *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, Korea, November.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), "Attention is all you need", *Proceedings of Advances in Neural Information Processing Systems*, Long Beach, CA, USA, December.

Wang, K., Zhang, D., Li, Y., Zhang, R. and Lin, L. (2017), "Cost-Effective active learning for deep image classification", *IEEE Transact. Circuits Syst. Video Technol.*, **27**(12), 2591-2600 https://doi.org/10.1109/TCSVT.2016.2589879

Wang, X., Girshick, R., Gupta, A. and He, K. (2018), "Non-local neural networks", *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, June.

Wang, Z., Xu, G., Ding, Y., Wu, B. and Lu, G. (2020), "A vision-based active learning convolutional neural network model for concrete surface crack detection", *Adv. Struct. Eng.*, **23**(13), 2952-2964. https://doi.org/10.1177/1369433220924792

Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K. (2017), "Aggregated residual transformations for deep neural networks", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July.

Xie, S., Feng, Z., Chen, Y., Sun, S., Ma, C. and Song, M. (2020), "DEAL: Difficulty-aware Active Learning for Semantic Segmentation", *Proceedings of the Asian Conference on Computer Vision*, Kyoto, Japan, December.

Xu, Y., Bao, Y., Chen, J., Zuo, W. and Li, H. (2019), "Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images", *Struct. Health Monitor.*, **18**(3), 653-674. https://doi.org/10.1177/1475921718764873

Yang, L., Zhang, Y., Chen, J., Zhang, S. and Chen, D.Z. (2017), "Suggestive annotation: A deep active learning framework for biomedical image segmentation", *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, Toronto, Canada, September.

Yang, X., Li, H., Yu, Y., Luo, X., Huang, T. and Yang, X. (2018), "Automatic pixel-level crack detection and measurement using fully convolutional network", *Comput.-Aided Civil Infrastr. Eng.*, **33**(12), 1090-1109. https://doi.org/10.1111/mice.12412

Ye, X.W., Jin, T. and Yun, C.B. (2019), "A review on deep learning-based structural health monitoring of civil infrastructures", *Smart Struct. Syst., Int. J.*, **24**(5), 567-586. https://doi.org/10.12989/sss.2019.24.5.567

Yoo, D. and Kweon, I.S. (2019), "Learning loss for active learning", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June.

Zhang, Y., Gao, J. and Zhou, H. (2012), "ImageNet classification with deep convolutional neural networks", *Adv. Neural Inform. Process. Syst.*, **25**, 1097-1105.

Zhang, X., Rajan, D. and Story, B. (2019), "Concrete crack detection using context-aware deep semantic segmentation network", *Comput.-Aided Civil Infrastr. Eng.*, **34**(11), 951-971. https://doi.org/10.1111/mice.12477

Zhao, H., Shi, J., Qi, X., Wang, X. and Jia, J. (2017), "Pyramid scene parsing network", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July.

Zhou, C., Bai, J., Song, J., Liu, X., Zhao, Z., Chen, X. and Gao, J. (2018), "ATRank: An attention-based user behavior modeling framework for recommendation", *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, February.