Measurement 217 (2023) 113091

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Semi-supervised learning framework for crack segmentation based on contrastive learning and cross pseudo supervision

Chao Xiang [a], Vincent J.L. Gan [b], Jingjing Guo [c], Lu Deng [d,*]

[a] College of Civil Engineering, Hunan Univ., Changsha, China
[b] Department of the Built Environment, National University of Singapore, Singapore
[c] College of Civil Engineering, Hunan Univ., Changsha, China
[d] College of Civil Engineering, Key Laboratory of Damage Diagnosis for Engineering Structures of Hunan Province, Hunan Univ., Changsha, China

ARTICLE INFO

Keywords:
Crack segmentation
Semi-supervised
Contrastive learning
Cross pseudo supervision
Multiple types of cracks

ABSTRACT

Fast and accurate crack segmentation plays an important role in the predictive maintenance of constructed facilities and civil infrastructures. However, it is worth noting that current deep-learning-based algorithms for crack segmentation may face significant challenges due to the requirement of a large amount of labeled data for high-precision segmentation. A novel semi-supervised learning framework for crack segmentation, which is referred to as semi-supervised crack (SemiCrack), based on the combination of contrastive learning and cross pseudo supervision (CPS) is presented in this study. The proposed segmentation network, called transformer and convolutional network (TC-Net), has a novel parallel encoder that fuses a transformer and a convolutional neural network. The inclusion of CPS can force the two models to maintain consistent outputs for various perturbed data based on the similarity loss. To capture the feature differences between positive and negative sample pairs extracted by the classifier and projector, pixel contrastive loss was also proposed. Compared with many state-of-the-art fully-supervised and semi-supervised segmentation algorithms, the results show that SemiCrack performs best on various publicly available datasets. The segmentation accuracy of TC-Net is higher than that of most fully-supervised segmentation networks, with an improvement of about 2% in Intersection of Union (IoU). Besides, SemiCrack requires only 20% labeled data to achieve comparable accuracy to other fully-supervised algorithms that require 100% labeled data. When the amount of labeled data is small, the IoUs of SemiCrack are significantly improved compared to fully supervised and semi-supervised networks.

1. Introduction

Crack detection is essential for ensuring the safety and functionality of civil infrastructures [1]. Crack-image segmentation is a technique that enables the accurate identification and extraction of crack pixels from non-crack pixels in an image [2]. As a result, it provides an essential quantitative tool for crack detection, damage assessment, and structural health diagnosis [3]. Due to the complex detection environment, crack segmentation methods based on conventional image processing not only require intensive manual intervention for feature identification and extraction, but also have poor generalization performance [4]. Many deep-learning-based crack segmentation models have been proposed as a result of the rapid advancement of deep-learning algorithms [5], including encoder-decoder-based networks [6] and attention-based networks [7]. To further improve the detection accuracy and efficiency, researchers have made the following modifications: (1) applying deeper pre-trained models as encoders [8]; (2) building new decoders containing more information [9]; (3) designing new feature fusion modules to improve the extraction of multi-scale information [10,11]; (4) changing the connection mechanism to enhance the ability to capture global contextual information [12]; (5) proposing new loss functions to improve the accuracy of model training [13]; and (6) incorporating robust pre-processing and post-processing methods [14,15].

Recent research has shown that a relatively large crack dataset with labels is essential for the high-quality performance of existing fully-supervised segmentation algorithms [16]. However, labeling cracks in image pixels can be challenging, particularly when the cracks are small

* Corresponding author at: College of Civil Engineering, Key Laboratory of Damage Diagnosis for Engineering Structures of Hunan Province, Hunan Univ., Changsha, China.
E-mail addresses: xiangchao@hnu.edu.cn (C. Xiang), vincent.gan@nus.edu.sg (V.J.L. Gan), guojingjing@hnu.edu.cn (J. Guo), denglu@hnu.edu.cn (L. Deng).

https://doi.org/10.1016/j.measurement.2023.113091
Received 30 December 2022; Received in revised form 18 May 2023; Accepted 20 May 2023
Available online 25 May 2023

or difficult to see with the naked eye [17]. Firstly, the complexity of crack shapes and the uncertainty of boundaries make it time-consuming to annotate the crack pixels with precision [18]. Secondly, the appearance and patterns of cracks vary greatly with the environment and materials of the target structure, making it unreasonable to reuse a crack dataset in different scenarios and necessitating new crack dataset labeling for each detection scenario [19]. In summary, the process of manually labeling crack pixels in a crack image dataset is extremely time-consuming and labor-intensive, making it expensive to create a large and accurately labeled crack dataset [20,21]. As such, the vast majority of available crack segmentation datasets are small in size and frequently overfit in fully supervised training. However, in practical crack detection, unlabeled crack data are readily available [22]. As a result, building a deep learning framework that can accurately segment cracks using a small number of labeled samples and a large number of unlabeled samples has gained attention from many researchers.

To avoid the need for a significant amount of labeled data, some unsupervised learning techniques [23,24] have been proposed. However, their accuracy is insufficient for the efficient segmentation of crack images. In contrast, the semi-supervised learning (SSL) approach enables the design of crack segmentation models with just a few labeled samples. It can strike a better balance between fully-supervised learning and unsupervised learning [25,26]. SSL-based segmentation models are trained using a small number of images with pixel-level labels and a sizable number of available unlabeled images. This drastically reduces the labeling effort while maintaining accuracy by making full use of unlabeled data. This approach effectively accelerates the development of segmentation models for practical applications [27].

SSL segmentation algorithms have been gaining increasing attention, as they can effectively train robust models for crack segmentation by extracting more useful information from unlabeled data. Several SSL algorithms for image segmentation have been proposed, such as Mean Teacher (MT) [28], Entropy Minimization (EM) [29], Deep Adversarial Networks (DAN) [30], Uncertainty Aware Mean Teacher (UAMT) [31], Interpolation Consistency Training (ICT) [32], Uncertainty Rectified Pyramid Consistency (URPC) [33], and Cross Pseudo Supervision (CPS) [34]. In practice, these algorithms can be categorized into two groups: (1) learning based on pseudo-labels and (2) learning based on consistency. Pseudo-labels are predicted for unlabeled images and combined with real-label images to increase the amount of training data. However, the quality of pseudo-labels can vary and impact the segmentation results obtained from this self-training-based approach. Consistency learning is a method that involves using two deep learning models to process the same unlabeled image and verify the consistency of their inference results when applying various perturbations, such as horizontal flipping, different contrasts, and brightness levels. This approach fully leverages the information provided by unlabeled data and improves the model's robustness and generalization capabilities by comparing the output results of the models [35]. The SLL-based approach discussed above primarily relies on the intrinsic properties of the dataset distribution rather than the assumptions of individual images. Therefore, a large number of unlabeled images as opposed to just labeled images are used for parameter optimization in training.

Currently, target detection and image classification in the field of crack detection use SSL algorithms. For instance, Guo et al. [36] proposed a semi-supervised detection framework for surface defect classification based on the mean-teacher algorithm. It can increase classification accuracy by more than 5% in comparison to the fully-supervised model when only 10% of the data in the dataset was labeled. He et al. [37] proposed a new SSL algorithm for classifying steel surface defects based on generative adversarial networks with an accuracy improvement of 16% compared to the previous algorithm. By including an attention mechanism in SSL, Karaaslan et al. [38] achieved rapid detection of cracks and spalling with a limited training set. The combination of SSL algorithms with generative adversarial networks was suggested by Liu et al. [39] as a way to classify crack images with high accuracy while retaining a low level of labeled data. Shim et al. [19,40] and Li et al. [25] proposed a semi-supervised segmentation framework based on adversarial learning to improve the crack segmentation accuracy of concrete and pavement images. They augment the dataset by using a generator to obtain pseudo-labels for unlabeled data, and then combine the pseudo-labeled data and the true-labeled data to train a powerful discriminator. However, it is frequently challenging to train generative adversarial networks for segmentation with the optimal accuracy. With the help of the base segmentation network powered by EfficientUNet, Wang et al. [22] implemented a semi-supervised crack segmentation based on the mean-teacher algorithm, which can effectively extract crack features. In addition, the interaction of student and teacher models can greatly reduce the amount of data needed for models to achieve the highest level of accuracy. However, the accuracy of the method for microcrack detection in complex backgrounds still needs to be improved [41]. With little focus on crack segmentation, the majority of current SSL algorithms are concerned with crack target detection. Although unlabeled images can be used by the semi-supervised algorithms mentioned above to help improve the segmentation performance. However, the loss of global contextual information due to repeated downsampling may still make it difficult to improve segmentation accuracy [42]. The creation of a robust deep learning model is likewise significantly influenced by the size and diversity of the training set. The impact of the training set size on the performance of the crack segmentation model in various detection scenarios has received little attention in the preceding studies.

Contrastive learning [43] and CPS [34] are two recently proposed SSL methods that are known to be effective in improving the performance of SSL models [44]. However, while CPS with contrastive learning has demonstrated potential in SSL segmentation, the use of these two algorithms in crack segmentation needs further investigation. Moreover, current SSL networks for crack segmentation rely on convolutional operations, which are not suitable for extracting global information from the background. As a result, most SSL segmentation algorithms have poor performance in segmenting crack images with complex backgrounds [45]. The transformer, a popular self-attention technique, has been shown to improve global information extraction and segmentation accuracy in segmentation tasks [46–49].

To solve the above challenges, this study aims to establish a new semi-supervised segmentation framework for crack segmentation (referred to as SemiCrack) based on contrastive learning and CPS. Besides, a new crack segmentation network fusing transformer and convolutional neural network (CNN), called transformer and convolutional network (TC-Net), is proposed to enhance the performance of the semi-supervised crack segmentation. The proposed framework outperforms other existing semi-supervised algorithms by achieving higher accuracy with only a small amount of labeled data, making it more practical in real-world engineering applications. This study contributes to new algorithms and findings in the following aspects.

(1) The proposed encoder of TC-Net integrates the strengths of two distinct structures, where the CNN emphasizes the local information of cracks, while the transformer models the long-range dependence of backgrounds. This significantly improves model accuracy in segmenting crack images with complex backgrounds.
(2) By incorporating CPS, the training process of SemiCrack fully combines the perturbations of the model and data, which expands the training data by using unlabeled data with pseudo-labels. Thus, the training quality of the segmentation network and the generalization ability of the model are improved.
(3) The integrated model with contrastive learning can learn from unlabeled samples and does not entirely rely on labeled samples for segmentation, saving time and manpower for tedious data labeling. While a small amount of labeled data is needed, the accuracy and robustness of the proposed new algorithms do not compromise. The developed contrastive loss function guarantees

the similarity of the learned knowledge of the two models, improving the accuracy and robustness of the model.

(4) A comprehensive experimental study is conducted using a variety of publicly accessible datasets to compare the proposed framework with other fully-supervised and SSL algorithms. The results demonstrate that the proposed segmentation network, TC-Net, outperforms most of the fully-supervised segmentation networks in the literature. With only 20% of labeled data, SemiCrack achieves 60.2%, 63.7%, and 70.6% of crack Intersection over Union (IoU) in the three datasets, respectively. It not only performs better than traditional advanced semi-supervised algorithms, but also achieves the accuracy of other fully-supervised algorithms with 100% labeled data using only 20% labeled data.

Following is the arrangement of the remaining sections. The specifics of the proposed method are described in Section 2, which also provides a brief of the loss function that is used. The datasets, training parameter settings, and evaluation metrics are the main topics of Section 3. The results and analysis of testing the proposed framework and other algorithms on various datasets are then presented in Section 4. Conclusions and future works are drawn in Section 5.

2. Proposed methodology

2.1. Overview

In real-world datasets, labeling crack images at the pixel level would be labor-intensive and time-consuming, whereas unlabeled crack images are easily accessible. An SSL framework based on contrastive learning and CPS is proposed for crack segmentation to learn richer and more useful information from unlabeled data. This framework takes into account the loss associated with a large amount of unlabeled data when calculating the target training loss. Thus, it can significantly improve the ability of models to generalize unknown data. Fig. 1 shows the training flow of the developed SemiCrack framework, which is made up of four main parts: segmentation networks based on the transformer and CNN (TC-Net1 and TC-Net2), classifier, projector, and combined loss function. The framework can effectively reduce label dependency by using rich unlabeled data along with limited labeled data for joint training. The proposed framework code is available at https://github.com/ChaoXiang661/SemiCrack/ for academic use.

The N labeled images (X_n) (whose label set is Y_n) and the M unlabeled images (X_m) (M≫N) make up the training set for general SSL. During the training process, all input samples ($X_n \cup X_m$) are added to the framework with varying degrees of perturbation for data enhancement. The enhanced labeled data (X_n) and unlabeled data (X_m) are then sent to TC-Net1 and TC-Net2. After each training iteration, the two segmentation networks output their respective predictions independently of one another with different parameter initializations. For further contrastive learning, their predictions are then fed to the classifier (for labeled data) or the projector (for unlabeled data). For prediction in the inference stage, crack images can be inputted into either of the segmentation network models to obtain segmentation results, without the involvement of the classifier and projector. In this study, a new loss function made up of supervised segmentation loss, similarity loss, and feature consistency loss is proposed to improve the convergence of models. The overall loss is calculated using the provided labeled and unlabeled data to train and update the model parameters of this framework.

To be more precise, X_n is first split evenly into the two disjoint subsets X_{n1} and X_{n2} ($X_{n1} \cup X_{n2} = X_n$). To obtain the crack probability distribution prediction maps P1n ($P1_{n1} \cup P1_{n2}$) and $P2_n$ ($P2_{n1} \cup P2_{n2}$), all of them with data enhancement are input to TC-Net1 and TC-Net2. After computing $P1_n$ and $P2_n$ using the sigmoid activation function, the corresponding crack segmentation maps S1n and S2n are then obtained. The differences between $S1_n$, $S2_n$, and Y_n, respectively, are used to calculate the supervised segmentation losses of the two networks. To

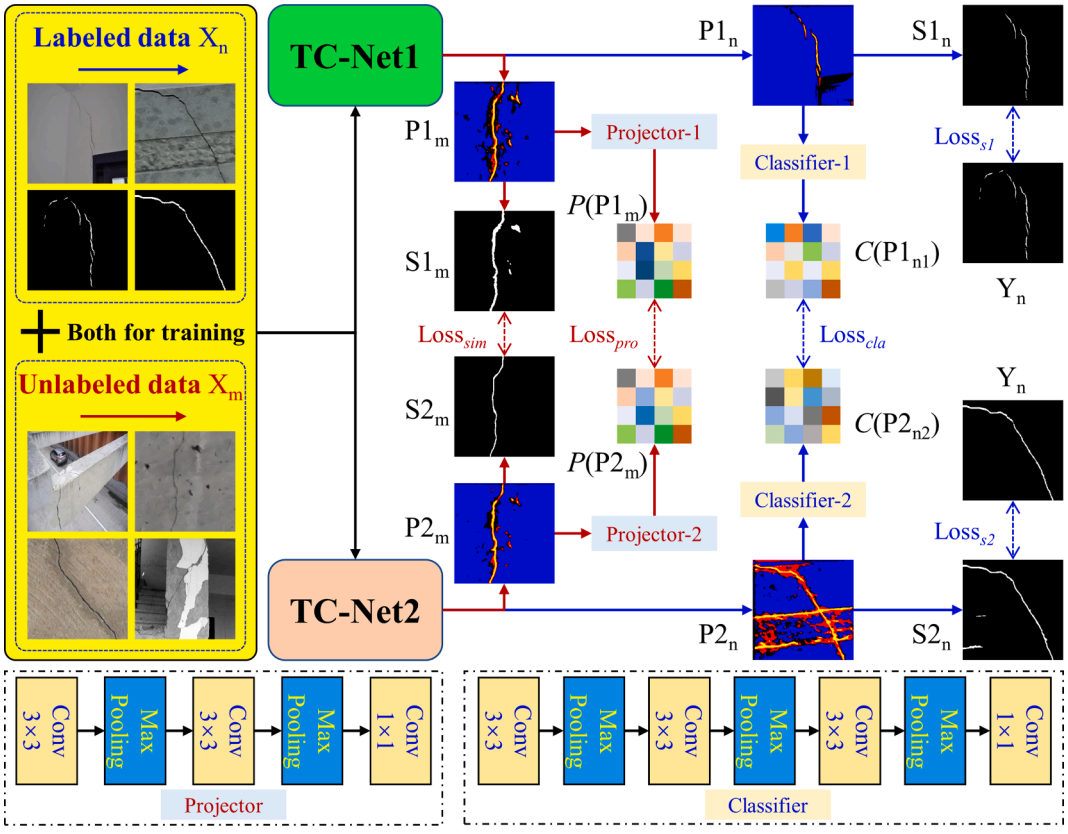

Fig. 1. Training flow chart of the proposed SSL framework (SemiCrack).

extract the feature representations $C(P1_{n1})$ and $C(P2_{n2})$ at the same spatial location for labeled data that is not the same, $P1_{n1}$ and $P2_{n2}$ are then input to the classifier module. And pixel-wise contrastive loss is applied to maximize the feature differences between $C(P1_{n1})$ and $C(P2_{n2})$.

X_m is fed into different segmentation networks after data enhancement to produce the corresponding crack probability distribution prediction maps P1m and P2m. The crack segmentation maps S1m and S2m are then calculated from the obtained $P1_m$ and $P2_m$ using the sigmoid activation function. Then $S1_m$ and $S2_m$ are compared to determine the similarity loss of both. The feature representations $P(P1_m)$ and $P(P2_m)$ of all unlabeled images are then extracted using $P1_m$ and $P2_m$ as inputs to the projector module. Finally, the proposed pixel-wise contrastive loss maximizes the feature differences for non-corresponding spatial locations while minimizing the feature differences for corresponding spatial locations between $P(P1_m)$ and $P(P2_m)$. The methodology details are presented in the following subsections.

2.2. Segmentation network based on transformer and CNN

Convolutional operations used in CNN segmentation models are good at extracting local features but provide limited support for extracting global representations. This restricts the performance improvement of CNN models in challenging segmentation tasks. A deep learning network structure that differs from CNN is a transformer, which consists of embedding and self-attention [50]. Recent studies have shown how well the transformer performs when it comes to extracting global information and modeling long-distance dependencies.

It has been demonstrated that the dual encoding structure can enhance the learning of features to achieve fine segmentation of images in complex environments [46,51,52]. A new deep learning segmentation network is proposed in this study based on the transformer and CNN (TC-Net), whose encoder is a combination of the CNN and transformer for crack segmentation. TC-Net not only maintains accurate local feature extraction, but also successfully models the global crack image. As shown in Fig. 2, the TC-Net involves six components including a transformer coding branch, CNN coding branch, scale-aware pyramid fusion module (SAPFM), feature fusion module (FFM), feature connection module (FCM) between codecs, and feature decoder. The need for a lot of data in the transformer network can be effectively reduced by combining the CNN and transformer in the encoder. Feature fusion in training can effectively take advantage of the advantages of both models to achieve better learning outcomes. The six components that constitute TC-Net are elaborated as follows.

2.2.1. CNN coding branch

A powerful, fine-grained feature extractor that can recognize distinctive local crack features is available from CNN. The proposed network uses a modified pre-trained ResNet-34 model [53] as one of the feature encoders, referring to TransFuse [46] and FAT-Net [52]. As depicted in Fig. 2, it has five convolutional blocks. Each convolutional block is made up of a convolutional layer, a batch normalization layer, and an activation function layer of rectified linear unit (ReLU). Additionally, the gradient vanishing problem that exists in deep neural networks can be effectively solved by using residual connections between each convolutional block. This will simultaneously hasten the network convergence. The CNN coding branch gradually extracts local semantic information from the original crack image using progressive downsampling.

2.2.2. SAPFM

To capture multi-scale contextual information, SAPFM is inserted on top of the encoder, which can encode the high-level semantic feature maps. As can be seen from Fig. 3, the proposed SAPFM [54] consists of three parallel atrous convolutional filters with shared weights and two cascaded scale-aware modules with spatial attention mechanisms. The number of model parameters and the risk of network overfitting can both be significantly decreased by using shared weights among the filters. The scale-aware module can create adaptive learning and feature fusion by introducing the spatial attention mechanism. Specifically, the weighted formula for the fused feature map is:

$$F_{fusion} = A \odot F_A + B \odot F_B \tag{1}$$

where A and B are pixel-wise attention maps created using the respective feature mappings. F_A and F_B stand for features at two different scales. And $\odot$ denotes the element product operation between the attention map and the feature. The fused feature maps of the three branches are fed into the two cascaded scale-aware modules to produce the final fused features. To obtain the output of the entire SAPFM, the feature maps before and after the input are finally connected by the residuals and the learnable parameter α. Through self-learning, SAPFM can dynamically choose the most suitable receptive fields for targets at different scales. Finally, it achieves the fusion of multi-scale contextual information and the improvement of crack segmentation accuracy.

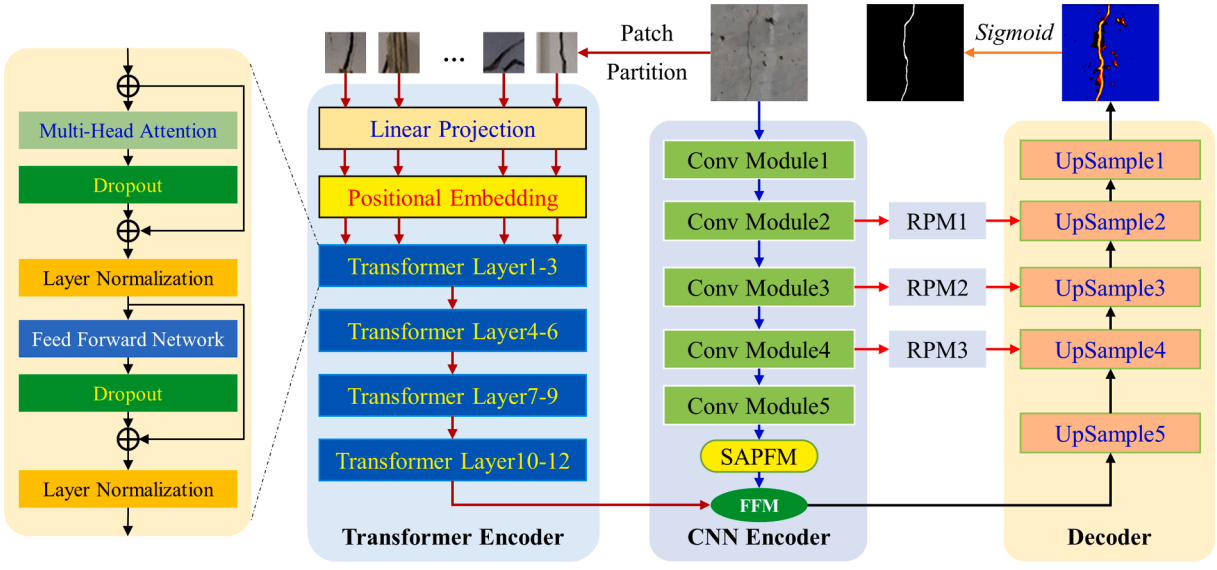

Fig. 2. The structure of the proposed segmentation network.

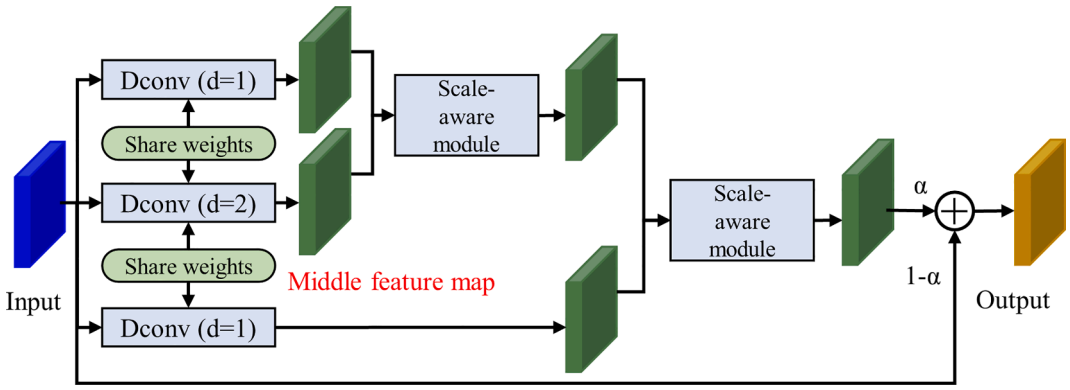

Fig. 3. The structure of the proposed SAPFM.

2.2.3. Transformer coding branch

The extraction of global features from crack images is achieved by the transformer, which is entirely dependent on the attention mechanism to model the global dependencies of input and output [50]. The input module and the encoding module make up the transformer encoding branch built for TC-Net [52].

Patch segmentation and position embedding operations make up the input module. The input module changes the input format [H, W, C] into a vector sequence that is compatible with the input format of the standard transformer module. The crack image is first sliced into a series of small, non-overlapping patches. Each patch is then fed to the embedding layer after being linearly mapped into a one-dimensional vector. The relative position relationship between the patch features is established with the help of the position embedding operation. It makes sure the model can locate feature markers precisely when computing attention, making it simpler for the model to pick up new information.

The encoding module is made up of a succession of identical network layers. Multi-Headed Attention (MHA) and Feedforward Network (FFN) are two additional divisions that can be made for each layer [50]. First, feature dependencies between template regions and search regions in the global feature representation are captured by the self-attentive mechanism of each layer. The attention mechanism can also efficiently combine global information and improve discriminative features, which aids the network in more precise target localization. Then, residual connection and normalization operations are also implemented between layers. The residual connection can be thought of as the result of superimposing the layer output and initial input. To avoid overfitting during model training and to accelerate model convergence, a layer of normalization operation is applied to the residual connection result.

2.2.4. FFM and FCM

While the CNN encoding branch excels at extracting local geometric features, the transformer encoding branch excels at extracting global features. To fully utilize both capacities for feature extraction, as shown in Fig. 2, the FFM [52] is embedded for fusing the top outputs of the two branches. The proposed FFM can decouple local features and global features in parallel. Thus, it can fully exploit the high efficiency of CNN in extracting local features and the powerful capability of the transformer in modeling global features. The resolution of the final output feature maps of both coding branches should be set to the same size. This will ensure the accuracy of segmentation while keeping the computational cost as low as possible. To fill the semantic gap between the encoder and decoder, a new residual path module called FCM [52] is created between the CNN coding branch and the decoder for feature connection. It can automatically change the feature mapping distribution of the encoder and decoder to improve model training and crack segmentation accuracy. Each layer of residual connection has a different number of base blocks, with the base block of the FCM consisting of a 3 $\times$ 3 and a 1 $\times$ 1 convolutional block.

2.2.5. Feature decoder

For quick and effective recovery to the high-resolution feature maps, a decoder with five upsampling blocks is also set up, as depicted in Fig. 2. The fused feature map from the top output of the two-branch encoder is the input of the first upsampling block. Details and spatial information can be extracted from the encoder through RPM to guide the decoder to produce the final segmentation map. The other upsampling blocks are all linked to the corresponding CNN encoding block [55]. The final block outputs the crack probability distribution map, and the crack segmentation result is obtained by the convolutional layer and sigmoid function calculator. The decoder employed can effectively combine the local features extracted from the CNN branch and the global background information extracted from the Transformer branch [52]. The proposed TC-Net can increase the feature extraction capability and improve the crack segmentation accuracy when compared to the original U-Net and transformer segmentation network.

2.3. Contrastive learning and cross pseudo supervision

The proposed framework incorporates cross pseudo supervision (CPS) [34] with contrastive learning [43] in the training process to lessen the reliance on a large amount of labeled data. In addition, it reduces the over-reliance on contextual information, while improving the extraction of global information. Therefore, it improves the segmentation performance of the model in unknown domains.

CPS is achieved by controlling the results of two segmentation networks with the same structure but different parameter initializations. During the parameter updates of iterative training, the addition of data perturbations forces the output results of both models to be consistent for the same data [34]. Both networks simultaneously output their respective segmentation predictions, which can be used as a pseudo-label for the other network. Then, the pseudo-label is used as a supervisory signal for unlabeled data. The diversity of the dataset can be greatly increased by using unlabeled data with pseudo-labeling in this manner.

In the proposed framework, positive and negative sample pairs are used to train the model in contrastive learning. Positive sample pairs refer to pairs of samples that are similar, while negative sample pairs refer to pairs of samples that are dissimilar. For a given image, the resulting high-level features should be similar regardless of the feature transformation. High-level features should be distinct at different spatial locations while being similar at the same spatial location for the same image [43]. A basic principle of contrastive learning used in this study is that the high-level feature representations extracted from the output of the same image after two different segmentation models should have a high degree of similarity. At the same time, these high-level feature representations should be different for different images. An appropriate contrastive loss is proposed to express the dissimilarity of such features during training. Contrastive learning can introduce a large amount of

unlabeled data during training to minimize pixel-wise loss at the corresponding location.

During training, the projector and classifier run as two independent branches without any feedback between them. The purpose of both is to efficiently extract high-level features from low-level semantic data [43], while ensuring that feature representations are distinct for different images, but consistent for the same images through segmentation networks. The convolutional layer with a kernel size of 3 × 3 and the max-pooling layer with 2 × 2 filters are the same fundamental building blocks in both the classifier and the projector, with the classifier having three of each and the projector having two. Both are then connected to a convolutional layer with a kernel size of 1 × 1. By feeding each crack image into two segmentation models with a different initialization, all crack images are first transformed. The feature representations extracted using the projector from similar images should be identical, whereas the feature representations extracted using the classifier from different images should be noticeably different. To capture these feature similarities and differences, suitable contrastive loss functions are created.

Specifically, the projector is used on the unlabeled image after it has been transformed into two different probability distribution maps by two different segmentation networks. The high-level semantic information extracted by the projector should be consistent for two probability distribution maps of the same unlabeled image. When using two identical pairs of images, the projector is used to create both positive and negative sample pairs. The positive sample pairs contain two samples with the same feature located in the same position, while the negative sample pairs contain two samples with different features located in different positions. Iterative training can reduce contrastive loss at the pixel level. Thus, the differences between the predictions of two networks for the same image are the smallest at corresponding spatial locations and the largest at non-corresponding spatial locations. The classifier then operates on two non-overlapping subsets of the labeled images, which are input to two different segmentation networks for producing four subsets of probability distribution maps. Negative sample pairs (i.e., different images with different features) are created using the classifier. A suitable loss function is constructed and trained iteratively to maximize the feature differences between the different labeled images extracted by the classifier. Thereby, the segmentation framework is encouraged to learn from different views of the labeled data.

2.4. Loss function

Five different components make up the loss function used in this study: supervised training losses for two segmentation networks with labeled data ($Loss_{s1}$, $Loss_{s2}$), contrastive learning loss with labeled data ($Loss_{cla}$), pseudo-supervised training loss with unlabeled data ($Loss_{sim}$), and contrastive learning loss with unlabeled data ($Loss_{pro}$).

For the labeled data, the predictions of the two different models are compared with the true labels to obtain $Loss_{s1}$ and $Loss_{s2}$. The aforementioned loss values are calculated using a combined loss made up of the binary cross entropy loss (BCE) and dice coefficient loss functions (Dice). The following is the definition of the overall supervised training loss:

$$Loss_{stotal} = Loss_{s1} + Loss_{s2} = Loss_{BCE} + Loss_{Dice} \tag{2}$$

where the binary cross-entropy loss is calculated as follows.

$$Loss_{BCE} = -\frac{1}{n}\sum[y_i \lg p_i + (1-y_i)\lg(1-p_i)] \tag{3}$$

where the dice coefficient loss is calculated as follows.

$$Loss_{Dice} = 1 - \frac{\sum p_i y_i + \varepsilon}{\sum(p_i + y_i) + \varepsilon} - \frac{\sum(1-p_i)(1-y_i) + \varepsilon}{\sum(2 - p_i - y_i) + \varepsilon} \tag{4}$$

where n denotes the total number of image pixels, y_i denotes the true value of the i-th pixel, and p_i denotes the prediction probability of the i-th pixel. When y_i and p_i are too small, training instability due to excessive gradient variation is prevented by the constant ε.

The labeled data is split into two non-overlapping sub-datasets before being fed into two segmentation networks. Then two sets of different crack distribution maps are selected from the four sets that do not belong to the same crack subset. The feature representations of these two different sets of crack images are then extracted using a classifier. The differences between these two sets of distribution maps are calculated using the pixel-wise contrastive loss [43], where $Loss_{cla}$ is calculated as:

$$Loss_{cla} = -\log\frac{1}{\sum\exp(qk_-/\tau)} \tag{5}$$

where q and k denote the feature representations of the two sub-datasets obtained using the classifier, and qk_- represents a negative sample pair, that is, two feature representations at the same position with different meanings. τ is a constant.

The combined loss function includes both binary cross entropy and dice coefficient loss functions. The training loss $Loss_{sim}$ for unlabeled data is calculated by comparing the predictions of two different models.

All unlabeled data are input into two models to get two sets of crack distribution maps. These two sets are fed into the projector to get the high-level semantic feature representation of crack images. The same crack image is transformed using two different segmentation models. The high-level feature representations obtained should be consistent. Thus, the pixel-wise contrastive loss is used to determine their similarity. The $Loss_{pro}$ is calculated as

$$Loss_{cla} = -\log\frac{\exp(qk_+/\tau)}{\sum\exp(qk_-/\tau)} \tag{6}$$

where qk_+ represents a positive sample pair, i.e., two feature representations at the same position indicate the same meaning.

Overall, the SemiCrack framework for segmenting cracks learns from both labeled and unlabeled data. The model achieves its highest segmentation accuracy by iteratively reducing the total loss during training. The $Loss_{total}$ is determined by:

$$Loss_{total} = Loss_{s1} + Loss_{s2} + Loss_{sim} + \lambda(Loss_{cla} + Loss_{pro}) \tag{7}$$

where λ is a weighted value created using a Gaussian function, which varies with the number of training iterations. This value helps balance the weights assigned to segmentation and contrastive losses during training.

3. Datasets and implementation

3.1. Datasets

A comprehensive experimental study was conducted on three distinct types of crack datasets (i.e., concrete cracks, pavement cracks, and steel cracks) to demonstrate and validate the proposed new framework. The effectiveness of SemiCrack in segmenting cracks across various structures is fully demonstrated and validated. Concrete crack images were derived from the open-source concrete crack dataset [56], which consists of four base datasets with more complex crack image backgrounds and varied crack patterns. Various types of concrete buildings are included, with over 50 bridges from Hangzhou, China, a tunnel in Huzhou, and existing housing buildings in Harbin, as shown in Fig. 4(a). Pavement crack images were obtained from the Crack 500 dataset [57], which has a total of 500 pavement crack images with a size of 2000 × 1500 pixels. It was taken by Yang et al. using a cell phone at the main campus of Temple University. Each image was labeled at the pixel level, as shown in Fig. 4(b). Steel crack images were obtained from the steel box girders of cable-stayed bridges used in the first International Project Competition for SHM 2020 (IPC-SHM 2020) [58]. This

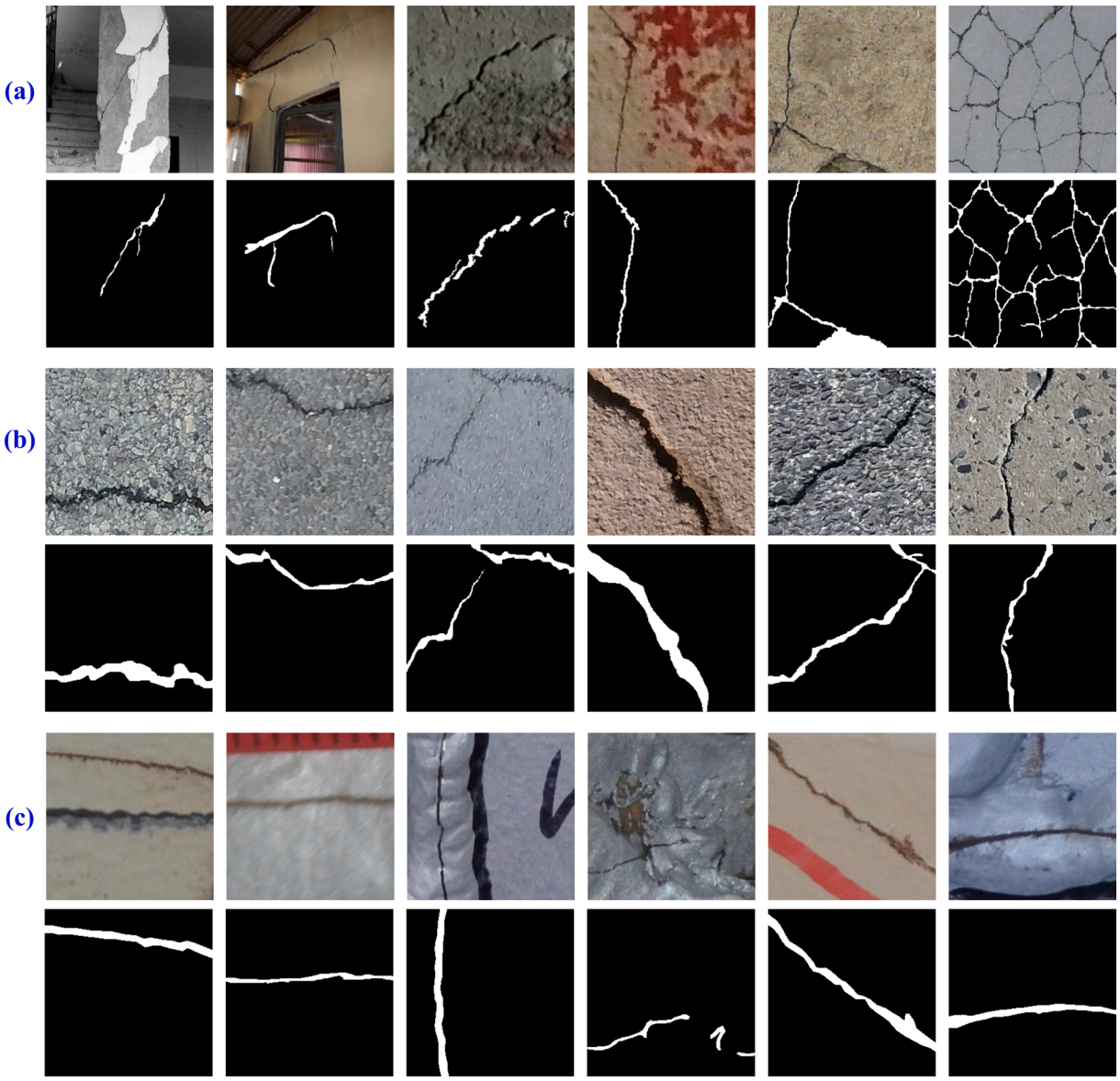

Fig. 4. Examples of crack images contained in the three datasets.

dataset was captured using a Nikon D7000 camera at a resolution of either 4928 × 3264 pixels or 5152 × 3864 pixels and contains more complex interference information, as shown in Fig. 4(c).

Due to the current hardware setup, all original images in the three datasets were cropped into the image with a size of 256 × 256 pixels for computation and model training. Additionally, only images with more than 1000 pixels of cracks were chosen to ensure that there are sufficient crack features available for learning. The final three datasets are presented in detail in Table 1 after being split into training, validation, and test sets in the ratio of 8:1:1.

Table 1
The details of the three datasets.

Crack dataset	Total number	Training	Validation	Testing
Concrete crack	3238	2590	324	324
Pavement crack	6890	5512	689	689
Steel crack	1773	178	178	1417

3.2. Implementation

All experiments were trained and implemented using the PyTorch on a server with an NVIDIA RTX 1080Ti GPU. The network weights were updated using the Stochastic Gradient Descent (SGD) algorithm with a momentum of 0.9 and a weight decay of 1e-4 to ensure the convergence of the network. The number of training epochs was 200 for fully-supervised training and 500 for semi-supervised training. If the model does not improve on the validation set for 50 consecutive epochs in semi-supervised training, the training is terminated using an early-stopping strategy. The initial learning rate was 0.001, and the learning rate was adjusted during training using a poly strategy, where the initial learning rate was multiplied by a power of 0.9 after each epoch.

During the training process of SemiCrack, a part of the training set was divided into labeled data according to the experimental set ratio, while the rest of the training set was divided into unlabeled data. The same validation set and test set were used for all experiments for fairness and objectivity of comparison. Numerous data augmentations were used during training, such as rotation, flipping, affine transformations, etc., to enhance the capacity to learn general representational features and to increase the diversity of image samples to avoid overfitting. Loss was

chosen as a monitoring indicator to guide the model to update parameters (weights and biases). After each training epoch, the F1-score of the validation set was calculated, and the model with the lowest F1-score was ultimately kept. Each set of experiments in this study was trained three times to avoid occasional peaks in segmentation accuracy. The final experimental results were then obtained by averaging to reduce the effect of random initialization and obtain more stable results. A dual-stream sampling strategy was used for semi-supervised training, loading labeled and unlabeled data in the same batch. The batch size for semi-supervised training was set to 16, which included eight labeled images and eight unlabeled images, due to the limited GPU memory and computational efficiency. The batch size for fully-supervised training was also set to 16.

3.3. Competing algorithms and evaluation metrics

Several comparative experiments were carried out to assess the effectiveness of the proposed framework. First, it is investigated how different crack segmentation models perform when training on three different types of crack datasets using different sizes of training sets. The proposed TC-Net was compared to other classical crack segmentation networks, including UNet-ResNet34 [59], DeepCrack [60], DcsNet [9], TransFuse [46], UTNet [61], and UCT [48] in fully supervised. The six networks were chosen because UNet-ResNet34, DeepCrack, and DcsNet are current CNN-based networks with high accuracy in crack segmentation, while TransFuse, UTNet, and UCT are high-precision segmentation networks based on a combination of the CNN and Transformer. The effectiveness of SemiCrack was then examined using three datasets with variously scaled training sets. Finally, the proposed method was compared with the most advanced semi-supervised algorithms on three datasets, including MT [28], DAN [30], UAMT [31], ICT [32], URPC [33], and CPS [34].

The widely used Intersection over Union (IoU) for the quantitative evaluation of crack segmentation is used to assess the effectiveness of the method. The proportion of background pixels is much higher than that of crack pixels for the whole dataset. To more accurately express the crack localization performance, only the crack IoU was calculated without taking the background IoU into account. The crack IoU is used to evaluate the size of overlap between the real crack region and the predicted crack region, and its calculation formula is

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{8}$$

where TP is the number of pixels that have been correctly identified as true cracks, FP is the number of pixels that are incorrectly identified as true cracks, and FN is the number of pixels that are mistakenly identified as non-cracks.

4. Results and discussion

In this section, the effect of the size of training sets on the prediction performance of the crack segmentation model in three different types of datasets was first investigated. The three datasets were then used to assess the proposed framework. Finally, SemiCrack was compared with various current state-of-the-art (SOTA) semi-supervised algorithms.

4.1. Effect of the size of the training set

The main purpose of this section is to confirm the impact of the training set size on model performance during fully-supervised training. Various sets of experiments were set up on three different types of datasets to compare model performance. Different sizes of labeled data were randomly selected as training sets in fully-supervised training, with proportions of 1%, 2%, 5%, 10%, 20%, 50%, and 100%, respectively, while other unextracted data were not used for training. Six current SOTA fully-supervised segmentation networks, including UNet-ResNet34, DeepCrack, DcsNet, TransFuse, UTNet, and UCT, were compared to the proposed TC-Net.

4.1.1. Concrete cracks

The corresponding segmentation models were first obtained after all networks had been trained on various numbers of concrete crack training sets. In Table 2, where the Mean row denotes the average of the results of the six existing networks available for comparison, the IoUs of the predicted results after testing on the same test set are displayed. The amounts of the labeled-data used in training are displayed in the various columns. Table 2 shows that when the proportion of labeled data is 1%, 2%, 5%, 10%, 20%, 50%, and 100%, respectively, during the fully-supervised training, the IoUs of TC-Net are 36.0%, 44.1%, 50.2%, 52.4%, 56.1%, 59.4%, and 62.1%, respectively. The IoU of TC-Net is about 2–3% higher than the average IoU of the other six networks at the same proportions. TC-Net performs significantly better than other networks with the same proportions of labeled data. In addition, the TC-Net model obtained by training on smaller datasets also outperforms some other networks on larger datasets. When the labeled proportion is increased from 1% to 100%, the IoU of TC-Net for the concrete crack dataset rises from 36.0% to 62.1%. Its IoU is increased by about 26%, and the percentage of increase is about 73%.

The comparison of the predictions made by TC-Net with different data amounts in concrete cracks is shown in Fig. 5. It can be seen from Fig. 5 that as the number of labeled data increases, the predictions become more accurate. The proposed network can extract the crack contours accurately in a variety of complex backgrounds when the proportion of labeled data is 100%. It demonstrates how TC-Net can successfully combine the benefits of a transformer and CNN. Besides, it also shows how the detection model works better with complex types of cracks. As seen in the first three rows of Fig. 5, some images in the concrete crack dataset have complicated backgrounds, and the size of the training set significantly affects the predictions of these images. The larger the training set is, the fewer cracks will be missed and incorrectly detected in the prediction. As shown in the fourth row of Fig. 5, it illustrates that the size of the training set in a simple background has less effect on the missed detection in the prediction.

4.1.2. Steel cracks

All networks were trained on different amounts of the steel crack dataset to obtain the detection models, and Table 3 shows the IoUs of different models for the test set. As can be seen from Table 3, TC-Net performs significantly better than the other segmentation networks in terms of IoU at any size of labeled data for the steel crack. The IoU of TC-Net at the same proportion is about 3% higher than the average IoU of the other six networks. When the proportion of labeled data in the training set is increased from 1% to 100%, the IoU of UNet-ResNet34 increases from 31.3% to 62.5% with a score increase of 31.2%, and its percentage increase is approximately 100%. The IoU of DcsNet increases from 35.0% to 63.4%, with an increase of 28.4% in the score, and its percentage of increase is 81%. The IoU of TransFuse increases from 37.5% to 63.7%, with a score increase of 26.2%, and its percentage of

Table 2
Concrete crack dataset.

IoU(%)	1%	2%	5%	10%	20%	50%	100%
Numbers	26	52	128	257	514	1285	2569
UNet-ResNet34	31.5	38.7	44.1	46.7	49.8	55.7	58.7
DeepCrack	30.6	39.3	44.0	47.0	50.2	55.3	58.3
DcsNet	33.4	41.8	46.8	49.9	53.1	57.5	60.1
UCT	34.4	40.6	45.6	50.6	54.3	57.9	59.9
TransFuse	34.8	42.1	47.5	49.7	55.3	58.1	59.5
UTNet	34.2	39.8	47.2	50.3	53.4	58.5	60.3
Mean	33.6	40.9	46.5	49.5	53.2	57.5	59.8
TC-Net	**36.0**	**44.1**	**50.2**	**52.4**	**56.1**	**59.4**	**62.1**

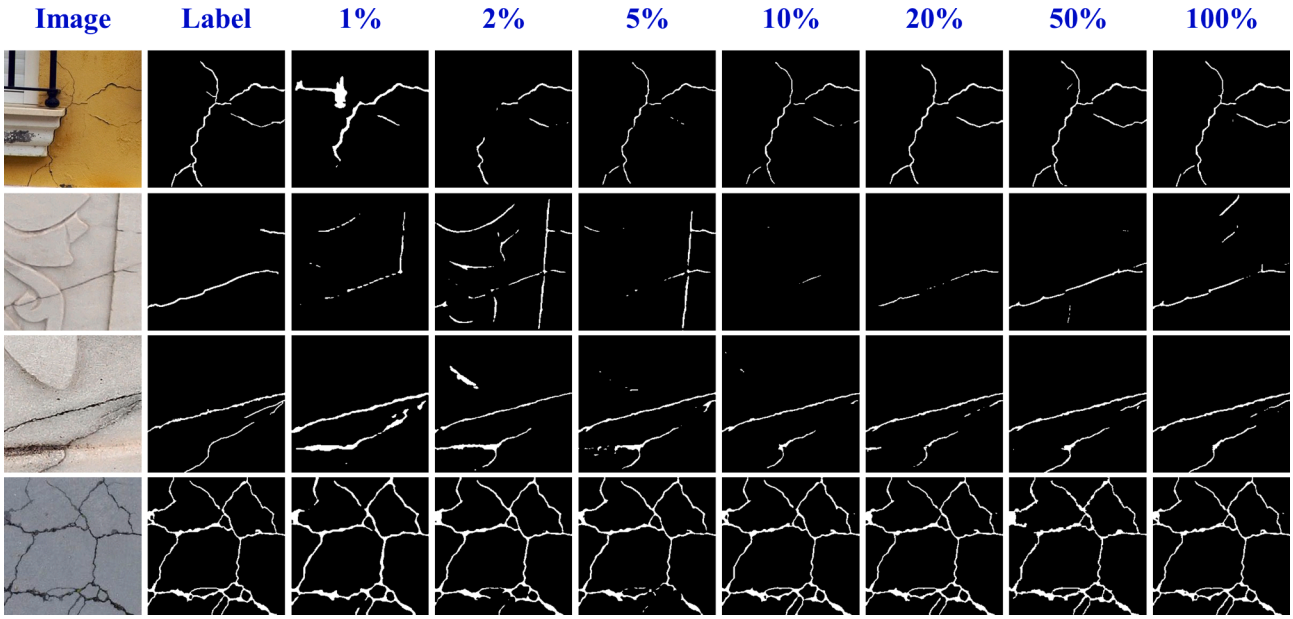

Fig. 5. Predictions of TC-Net with different data amounts in concrete cracks.

Table 3
Steel crack dataset.

IoU(%)	1%	2%	5%	10%	20%	50%	100%
Numbers	14	28	71	142	283	709	1417
UNet-ResNet34	31.3	35.1	48.5	52.4	55.2	57.8	62.5
DeepCrack	34.3	39.1	47.6	52.9	54.8	57.9	61.9
DcsNet	35.0	40.7	49.9	54.8	57.0	59.3	63.4
UCT	36.2	39.0	47.5	55.6	59.8	61.5	62.2
TransFuse	37.5	41.1	50.1	56.9	60.8	61.8	63.7
UTNet	37.3	40.8	49.0	56.6	58.6	62.0	63.9
Mean	35.9	39.9	49.2	55.3	58.2	60.5	63.4
TC-Net	**39.5**	**43.9**	**52.0**	**57.7**	**61.0**	**63.0**	**66.2**

increase is 70%. The IoU of TC-Net increases from 39.5% to 66.2%, with a score increase of 26.7%, and its percentage of increase is 68%. This indicates that the CNN-only-based network is more affected by the size of the training set in steel cracks than the network based on combining the transformer with CNN.

Fig. 6 shows the comparison for predictions of TC-Net with different data amounts on the steel crack dataset. As seen in the first three rows of Fig. 6, the steel crack dataset that is gathered in the field is highly influenced by elements like lighting, markings, and welding seams. Therefore, the identification is vulnerable to unclear positioning of the crack boundary. The trained model finds it difficult to extract crack features when the size of the training set is small. As a result, there are a lot of false detections and missed detections, which makes the prediction results very unsatisfactory. As the training set size grows, the false detection and missed detection rates of predictions significantly decline, and the shape of the predicted crack closely resembles the actual crack in the real label. As in the case of the fourth row of Fig. 6, when the training set is relatively small, the predicted value of crack width in the simple background is significantly larger than the true value. Therefore, the size of the training set at this time mainly affects the false detection rate of cracks and the crack width calculated from the selected boundary.

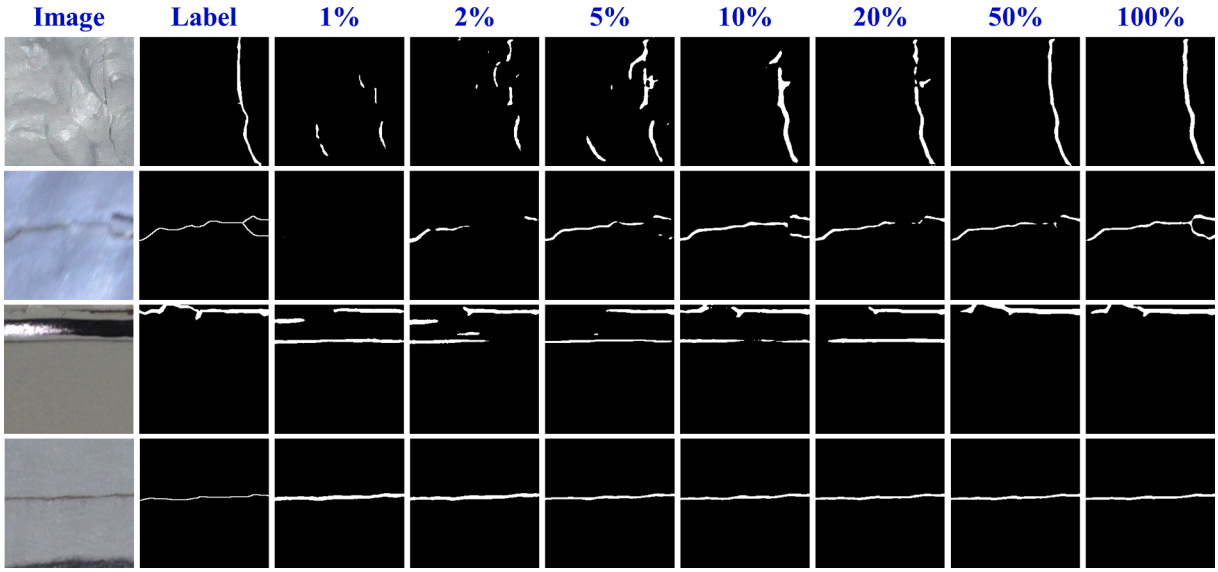

Fig. 6. Predictions of TC-Net with different data amounts in steel cracks.

4.1.3. Pavement cracks

The pavement crack dataset was used to train each network in varying amounts, and the obtained models were then used to predict the test set. Table 4 displays the IoU for each of the outcomes. The proposed TC-Net outperforms all other networks when the IoUs of all networks are compared in Table 4 at various amounts. When the proportion with labels gradually changes from 1% to 100%, the IoUs of TC-Net are 54.5%, 59.1%, 64.1%, 66.4%, 68.3%, 70.1%, and 71.8%, respectively. The IoU of TC-Net is about 1–2% higher than the average IoU of the other six networks at the same proportion. The increase in IoU for TC-Net and other networks is about 17% as a result of the gradual expansion of the size of the labeled data added to the training set. While the best IoU is only about 70% when 5512 images are used, the IoU of pavement cracks can reach more than 50% when only 55 images are used in training. It can be seen that the size of the training set has a relatively smaller impact on the IoU of the predictions for pavement cracks, increasing IoU by only about 35% compared to increases of more than 70% for concrete and steel cracks.

Fig. 7 shows the comparison for predictions of TC-Net with different data amounts on the pavement crack dataset. It can be seen from Fig. 7 that the predictions for pavement cracks with more complex shapes improve significantly with the increase of the data proportion. When the data proportion is 100%, the proposed TC-Net can effectively distinguish between crack and non-crack pixels in pavement images with different crack shapes. The proposed model can successfully distinguish between crack and non-crack features despite numerous competing factors in images.

4.1.4. Discussion

IoUs of all networks on the three types of crack datasets were analyzed based on the findings of the aforementioned three sections. Of all the networks, UNet-ResNet34, DeepCrack, and DcsNet are based on CNN, while UCT, TransFuse, UTNet, and the proposed TC-Net are based on a fusion of transformer and CNN architectures. The predictions of TC-Net are noticeably better than other networks on training sets of various sizes, especially when the size of labeled data is small. This validates the outstanding advantages of the proposed network on a small size of training data. When the proportion of labeled data is above 10%, UCT, TransFuse and UTNet have better IoUs for predictions than UNet-ResNet34, DeepCrack, and DcsNet. It also shows that the CNN-based alone network performs significantly worse at segmentation than UCT, TransFuse, and UTNet, which combine the benefits of both transformer and CNN structures. It is clear from the thorough analysis in Tables 2-4 that the IoUs of all networks significantly increase as more labeled data is added to the training set, but the gain gradually slows down. It demonstrates that the performance of both the CNN-based and transformer-based segmentation networks is significantly influenced by the quantity of training data. The decreasing increase indicates that the main factor affecting the model performance is not the amount of labeled data added to the training set when there is an adequate amount of labeled data available.

The extent to which the change in the training set size affects the model performance is also related to the size of the original training set. For example, in the TC-Net model, when the proportion of labeled data increases from 1% to 2%, the IoU of the three datasets increases by 8.1%, 4.4%, and 4.6%, respectively. When the proportion of labeled data increases from 50% to 100%, the IoU of the three datasets increases by 2.7%, 3.1%, and 1.7%, respectively. The percentage increase in IoU for all networks on the three different types of datasets as the proportion of labeled data increases rises 1% to 100% is shown in Fig. 8. Fig. 8 illustrates that when the size of the training set is small, increasing the size of the training set has a significant impact on the model performance. In comparison to UNet-ResNet34, DeepCrack, and DcsNet, the results of UCT, TransFuse, UTNet, and TC-Net, show a smaller percentage increase in IoU. This indicates that the CNN-based alone network is more influenced by the size of the training set than the fused Transformer and CNN-based network. Although UCT, TransFuse, and UTNet are also relatively less affected by the dataset size, their IoUs in the three datasets are lower than the proposed TC-Net. Additionally, it can be seen that the percentage increase in IoU for each network varies across the three datasets, being greater than 70% for concrete cracks, greater than 67% for steel cracks, and only slightly higher than 30% for pavement cracks. It demonstrates how the degree to which the size of the training set affects the model performance depends on the type of network and the type of structure in which the cracks are located.

Table 4
Pavement crack dataset.

IoU(%)	1%	2%	5%	10%	20%	50%	100%
Numbers	55	110	276	551	1102	2756	5512
UNet-ResNet34	51.1	54.9	60.4	65.0	66.9	67.9	68.9
DeepCrack	50.5	54.5	61.5	64.0	66.3	67.8	67.8
DcsNet	52.3	56.7	61.6	64.7	67.4	68.3	69.9
UCT	54.0	58.4	63.6	66.0	67.3	69.8	70.7
TransFuse	54.3	58.7	63.5	66.3	68.2	70.0	70.9
UTNet	53.8	58.5	63.0	66.0	67.9	69.0	70.5
Mean	53.0	57.3	62.5	65.5	67.5	69.0	70.1
TC-Net	**54.5**	**59.1**	**64.1**	**66.4**	**68.3**	**70.1**	**71.8**

4.2. Performance of SemiCrack

The effectiveness of the proposed SemiCrack and the impact of the amount of labeled data used in training on the functionality of the framework were assessed in this section. Each of the three crack datasets was used to train and validate SemiCrack. The proportions of labeled data selected during the model training were the same as in Section 4.1, which were 1%, 2%, 5%, 10%, 20%, and 50%, respectively. The performance variation of the model on the validation set during the training of the proposed framework is shown in Fig. 9. The monitoring metric used for the validation set is the F1-score, which takes into account both precision and recall.

The period of semi-supervised training was set to 500 epochs, and the early-stopping strategy was used for the training of all experiments. As can be seen from Fig. 9, the maximum number of training epochs is 370, proving that the parameter setting of the maximum training epoch is appropriate. The training status of the model significantly changes when different amounts of labeled data are added to the training set. When the proportion is higher, the detection model performs better on the validation set. Meanwhile, the amount of labeled data in training also affects the speed of model convergence and the number of training epochs. The model converges more quickly with more labeled data, which results in earlier training termination and fewer training iterations.

The optimal model was selected from the training of SemiCrack described above. The model performance was then evaluated on a test set and compared with the TC-Net model trained with full supervision. The IoUs of the SemiCrack models on the three different types of crack datasets are shown in Table 5. It should be noted that the model performance of SemiCrack improves with the increase in the proportion of labeled data for all datasets. The magnitude of performance improvement decreases with the presence of more labeled data. For concrete cracks, the requirement of a 20% labeled data results in 514 labeled images, achieving an IoU of 60.2%. Steel cracks require 283 labeled images, with an IoU of 63.7%. Pavement cracks demand 1102 labeled images, resulting in an IoU of 70.6%. In contrast, when all training data is labeled, the respective numbers of labeled images for concrete cracks, steel cracks, and pavement cracks are 2569, 1417, and 5512. Other fully supervised algorithms achieve the best IoU of 60.3%, 63.9%, and 70.9% on these three datasets, respectively. A very small difference (about 0.2%) separates the model performance of the proposed framework from that of other fully-supervised algorithms trained on the fully labeled training set. It indicates that the framework works well with training sets that have a large amount of unlabeled data and few labeled

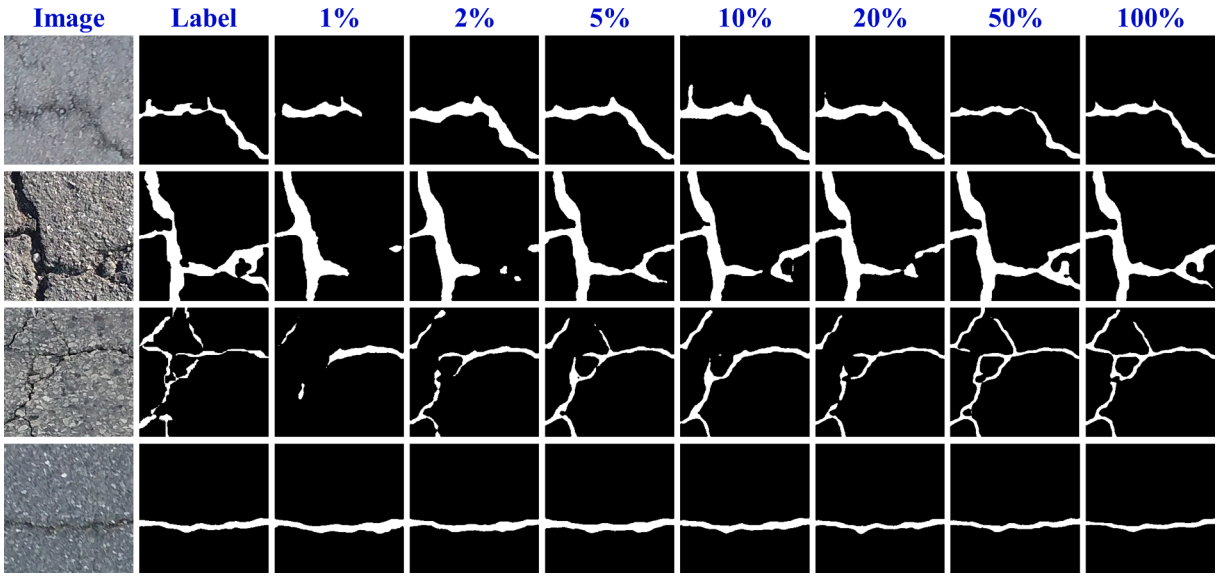

Fig. 7. Predictions of TC-Net with different data amounts in pavement cracks.

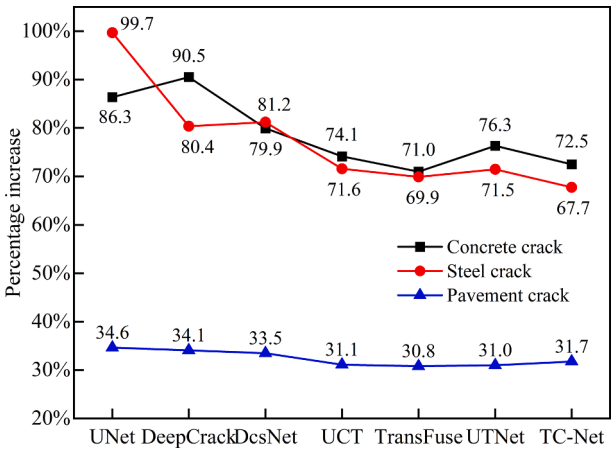

Fig. 8. Percentage increase in IoU with more labeled data (UNet means UNet-ResNet34).

data. The comparison results are sufficient proof that the framework effectively combines the advantages of contrastive learning and CPS as well as the effectiveness of the algorithm on various datasets.

The performance comparison of the proposed framework of SemiCrack with TC-Net (fully-supervised) on the test set under different conditions is shown in Fig. 10. As shown in Fig. 10, when there is a small amount of labeled data, the performance of the proposed framework is significantly better than that of TC-Net in fully supervised. As the proportion increases, the performance advantage of SemiCrack over the fully-supervised algorithm decreases. The proposed framework improves the fully-supervised algorithm by only 2% or less at a labeled proportion of 50%. The smaller performance improvement is caused by a combination of factors, including the upper bound for the optimal performance of the deep learning model on this type of dataset as well as variations in the amount of unlabeled data in training. For each kind of crack dataset, the overall number of images is fixed. During semi-supervised training, the number of images with labels rises while the number of unlabeled images sharply declines as the proportion of labeled data increases. For example, when the proportion of labeled data is 50%, the amount of labeled data is equal to the amount of unlabeled data. Semi-supervised training requires that the amount of unlabeled data is much larger than that of labeled data to fully exploit the performance of the semi-supervised algorithm. The results in Fig. 10 demonstrate that when the proportion of labeled crack data is 20% or less, the proposed framework performs significantly better than fully supervised algorithms. Therefore, in practical applications, if there is a certain amount of labeled data available, it is recommended to prepare at least five times more unlabeled data from the same source. This will fully leverage the advantages of the labeled data set and the performance of the proposed framework.

Some examples of predictions from SemiCrack on the three crack datasets are shown in Fig. 11. In Fig. 11, the first row shows the original crack images, the second row shows manual labels, the third and fourth rows show the predictions of the fully-supervised model of TC-Net at a proportion of 20% and 100%, respectively. The last row shows the predictions of SemiCrack at a proportion of 20%. As shown in Fig. 11, the fully-supervised trained model of TC-Net at 20% proportion is insensitive to crack pixels in more complex backgrounds and fails to detect insignificant cracks in images, where distracting factors in the background are mistakenly identified as cracks. In contrast, the proposed framework has a lower rate of missed and false detection and can better adapt to the challenging crack detection environment with the same data proportion. Additionally, the prediction maps show significant improvements in the shapes and boundaries of the cracks. On the other hand, SemiCrack uses a large number of unlabeled crack images in the model training. As a result, the model parameters can be updated to more accurately predict the low-resolution information in the crack images (as shown in the third column in Fig. 11).

4.3. Comparison with state-of-the-art algorithms

In this section, the proposed SemiCrack was compared with other current SOTA semi-supervised algorithms. Six of the most recent semi-supervised segmentation techniques, including MT, DAN, UAMT, URPC, CPS, and ICT, were chosen for comparison. Semi-supervised curvilinear (SemiCurv), an open-source SSL framework proposed by Xu et al. for curvilinear structure segmentation, is also included in the comparison. The fully-supervised TC-Net was trained on a labeled training set as the baseline model. These algorithms were re-implemented with the segmentation network changed from UNet to TC-Net. This lessens the impact of the segmentation networks used in the semi-supervised algorithms. Following that, the models were trained using training sets that contained varying amounts of labeled data. The IoUs of predictions for all models are displayed in Table 6 after they

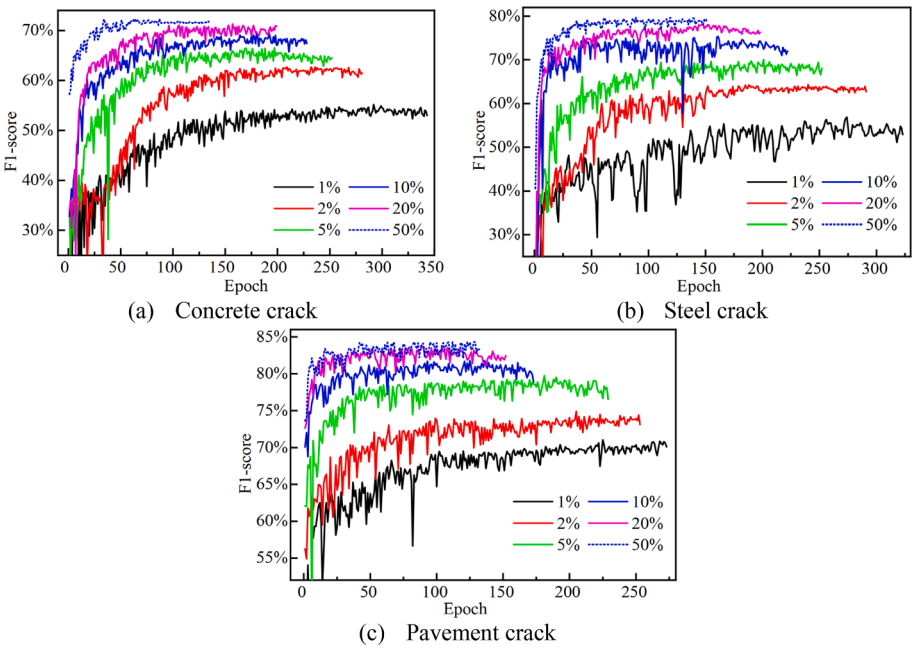

Fig. 9. Performance variation on the validation set during training.

Table 5
The IoUs of the SemiCrack models on the three types of crack datasets.

IoU(%)	1%	2%	5%	10%	20%	50%
Concrete crack	42.4	50.3	54.6	57.9	60.2	61.1
Steel crack	43.7	51.5	57.2	60.9	63.7	64.6
Pavement crack	58.1	61.4	66.1	68.3	70.6	71.0

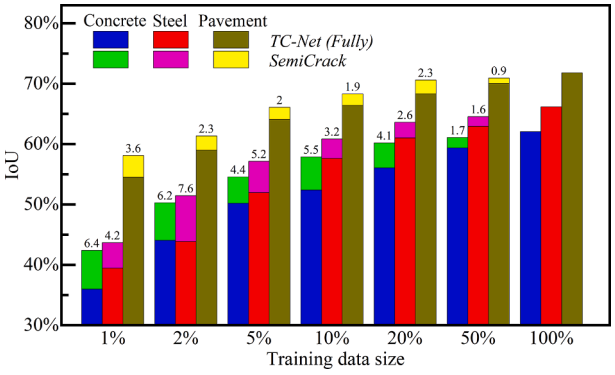

Fig. 10. Comparison of SemiCrack with fully-supervised TC-Net.

were tested on the concrete crack dataset.

The baseline model was obtained from fully-supervised training of TC-Net using just the corresponding amount of labeled data, so the results represent the lower bound of the segmentation results for the same amount. As shown in Table 6, all semi-supervised algorithms outperform the baseline model, indicating that they all benefit from the unlabeled data in training. Specifically, when the proportion of labeled images is 1%, the IoU of SemiCrack reaches 42.4%, which is 8.2% and 2.1% higher than that of the fully-supervised TC-Net and the best IoU in other semi-supervised algorithms, respectively. When the proportion of labeled images is 20%, the IoU of SemiCrack reaches 60.2%, which is 5.2% and 1.2% higher than that of TC-Net and the optimal semi-supervised algorithm, respectively. When the proportion of labeled data is less than 50%, the IoU of SemiCrack is improved by more than 5% over the baseline. As can be seen from Table 6, SemiCrack achieves the highest scores in all computational conditions. Its performance increases as the amount of labeled data for training increases. The segmentation performance of SemiCrack significantly exceeds that of fully-supervised algorithms and other semi-supervised algorithms.

Comparing the IoUs of the other semi-supervised algorithms, it can be found that among them the DAN algorithm has a lower IoU for all the amounts of labeled data. Concrete cracks come from a variety of image sources with complex backgrounds and varied crack structures. Thus, generative adversarial networks may be less successful at learning small target features like cracks. The addition of the contrastive learning algorithm is the main distinction between SemiCrack and the CPS algorithm. As demonstrated by the IoU improvement of the proposed framework over the CPS algorithm of about 3%, the improved portion of the proposed framework can effectively boost the performance of the crack segmentation model when the proportion of labeled data is less than 20%. Contrastive learning requires a large amount of unlabeled data to be fully beneficial as shown by the narrowing of the performance gap between the two when the amount of unlabeled data is not significantly greater than the amount of labeled data.

Three images from the concrete crack test set were chosen to further highlight the effectiveness of these comparative models. The outcomes of the proposed framework and other semi-supervised algorithms are shown in Fig. 12. It can be seen from Fig. 12 that the predictions of SemiCrack are closer to the real label than those of other models. SemiCrack can extract both the overall cracks in complex image backgrounds (the first row of Fig. 12) and the crack shapes at weak textures (the last two rows of Fig. 12). Based on the above analysis and visualization results, it can be seen that the proposed framework performs better for crack segmentation than other semi-supervised algorithms.

To show that SemiCrack performs better than other semi-supervised algorithms in two additional datasets with various kinds of cracks. The proportion of labeled data was assumed to be 20%. The IoU of each semi-supervised segmentation algorithm is shown in Table 7 following training. It is obvious from Table 7 that the proposed SemiCrack outperforms all other semi-supervised algorithms, demonstrating its broad applicability for crack detection in various scenarios. The proposed framework improves IoU on the steel crack dataset by 0.6% and on the pavement crack dataset by 0.4% when compared to other optimal semi-supervised algorithms.

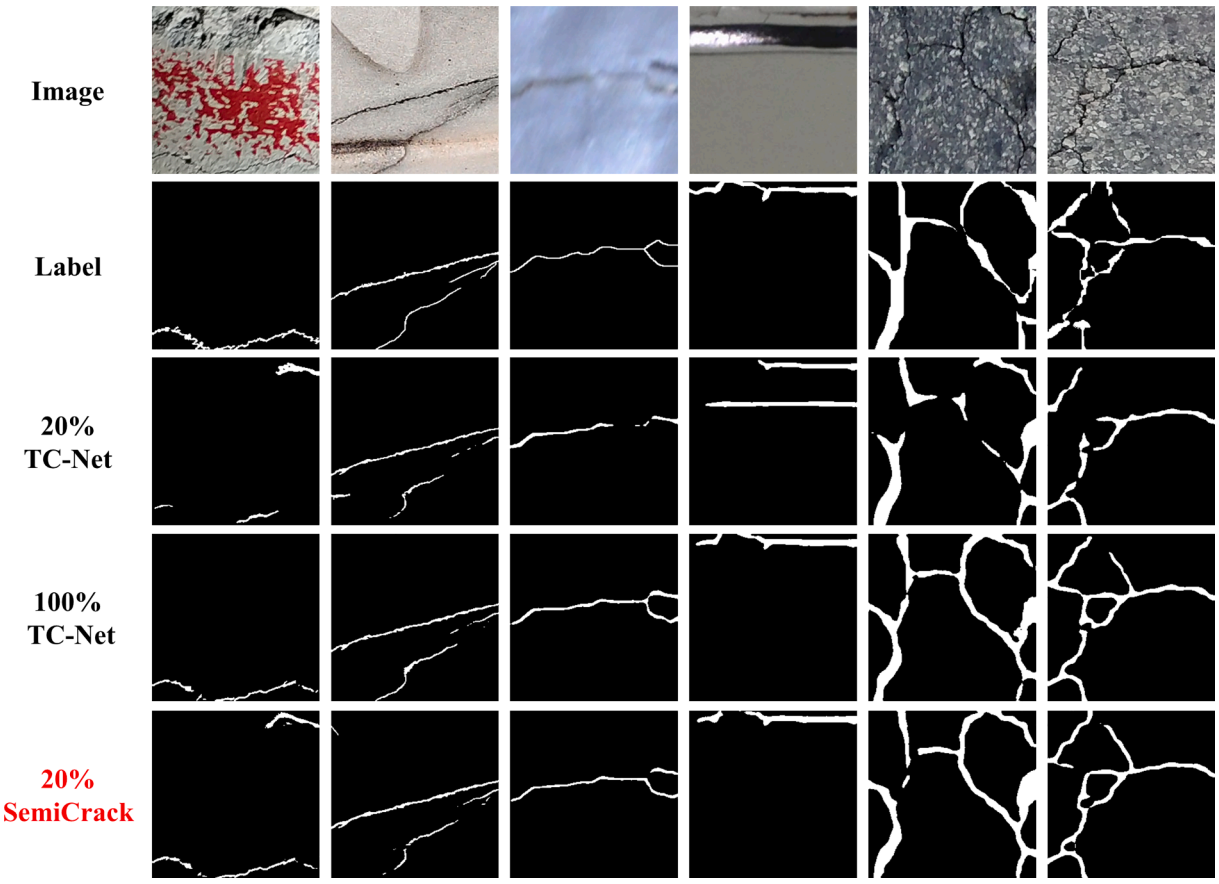

Fig. 11. Some prediction examples of the proposed SemiCrack and TC-Net.

Table 6
Comparison of the proposed framework with other advanced semi-supervised algorithms.

IoU(%)	1%	2%	5%	10%	20%	50%
Numbers	26	52	128	257	514	1285
Baseline	34.2	42.8	47.9	51.1	55.0	57.7
MT	39.8	46.1	50.5	55.0	58.6	59.5
DAN	36.3	43.5	48.3	52.2	55.7	57.9
UAMT	38.8	46.0	50.5	56.1	56.8	60.4
ICT	38.9	45.0	51.1	54.5	57.8	60.4
URPC	39.8	44.0	48.8	52.4	55.9	58.0
CPS	39.2	45.9	51.1	55.4	57.5	60.3
SemiCurv	40.3	47.0	52.2	56.6	59.0	60.6
Ours	**42.4**	**50.3**	**54.6**	**57.9**	**60.2**	**61.1**

4.4. Ablation experiments

Five experiments were conducted to demonstrate the effectiveness of each component in SemiCrack, including training TC-Net with cross pseudo supervision only (CPS), contrastive learning only (CL), removing the transformer branch (Transformer), removing the CNN branch (CNN), and removing SAPFM (SAPFM). Three types of crack datasets (concrete, steel, and pavement) with only 20% labeled data were used, and the IoU metric was used to evaluate detection accuracy. Parameter configurations were kept consistent across all experiments. The IoUs of the prediction results for all the experiments are shown in Fig. 13. Results show that SemiCrack outperforms other frameworks with an average improvement of 1.4% and 2.0% over CPS and contrastive learning only, respectively. The use of TC-Net with dual encoding branches and SAPFM results in an average improvement of 2.7%, 4.5%, and 0.9% compared to using a network without the transformer branch, CNN branch, and SAPFM, respectively, indicating their effectiveness in improving prediction accuracy.

5. Conclusion

Manual labeling of crack images is time-consuming and laborious, making unlabeled crack images easier to obtain than labeled ones. Hence, SemiCrack, a semi-supervised segmentation framework built on contrastive learning and cross pseudo supervision, is proposed in this study. It can effectively reduce the reliance on labels and learn crack features from more unlabeled crack images, which enhances model performance.

Several experiments were conducted on three crack datasets, including concrete structures, pavement structures, and steel structures, to validate the superiority of the proposed framework. First, how the model performance changes with the size of the training set during a fully-supervised training was investigated. Results show that the model performs better with more labeled data added to the training set. Then, the proposed TC-Net was compared with six current SOTA segmentation networks in fully supervised conditions. TC-Net outperforms the other segmentation networks in terms of IoU for any proportion of labeled data in the three different types of datasets. The IoUs of all networks improve significantly with the increase in the amount of labeled data added to the training set, but the gain gradually slows down. Additionally, SemiCrack was experimentally validated on three types of datasets. Results show that the framework with 20% of labeled data can achieve the best performance of other fully-supervised algorithms that require 100% of labeled data. Moreover, the experimental results show that SemiCrack achieved the highest scores in all computational conditions compared with other advanced semi-supervised algorithms. When the amount of labeled data is small, the IoU of the proposed

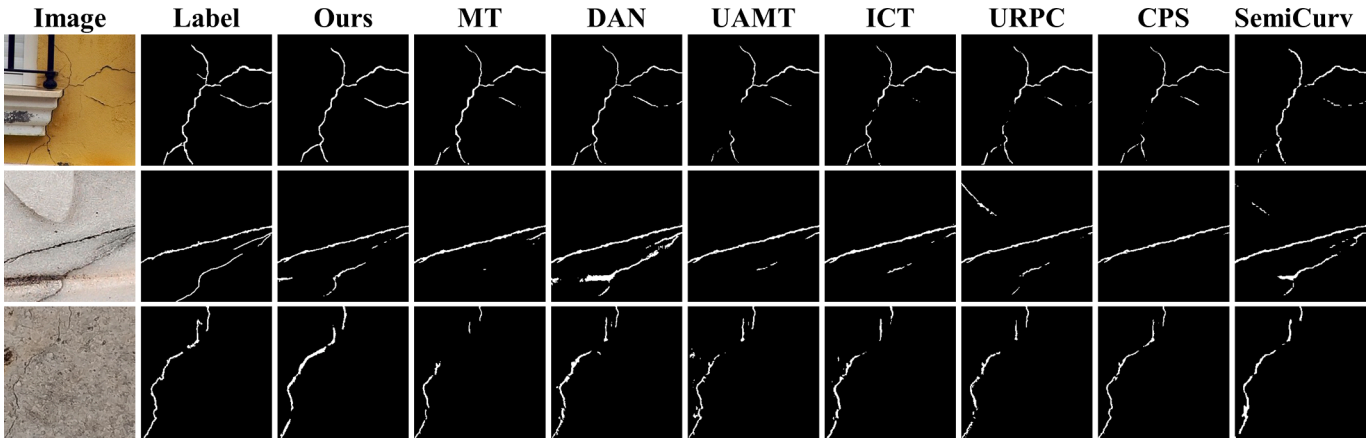

Fig. 12. Comparison of the proposed framework with other semi-supervised algorithms.

Table 7
Results for all semi-supervised algorithms at a proportion of 20% with labeled data.

IoU(%)	Baseline	MT	DAN	UAMT	ICT	URPC	CPS	SemiCurv	Ours
Steel	61.0	62.8	62.4	62.0	62.6	61.7	62.7	63.1	**63.7**
Pavement	68.3	70.1	69.4	69.3	69.8	69.1	70.0	70.2	**70.6**

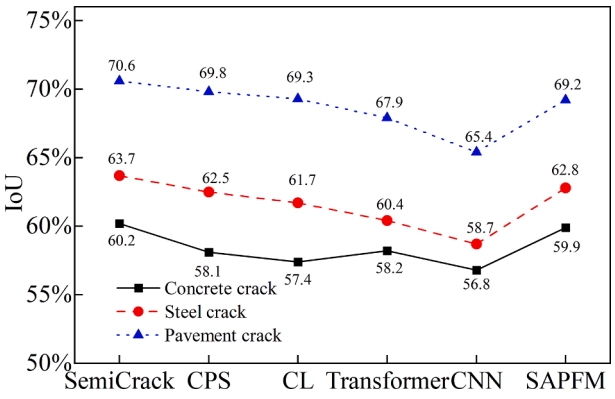

Fig. 13. IoU of results for the training framework with different components.

framework on the concrete dataset is improved by more than 2% over the optimal IoU of other semi-supervised algorithms.

SemiCrack incorporates a large number of unlabeled crack images to improve the performance of crack segmentation, but it still needs a specific number of crack images with precise labeling. In subsequent studies, algorithms will be developed to realize precise crack segmentation when only a few weakly labeled images are present.

CRediT authorship contribution statement

Chao Xiang: Formal analysis, Methodology, Data curation, Writing – original draft. **Vincent J.L. Gan:** Software, Validation. **Jingjing Guo:** Writing – review & editing. **Lu Deng:** Resources, Conceptualization, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The authors would like to acknowledge the financial support provided by the National Natural Science Foundation of China (Grant No. 52278177), the Hunan Province funding for leading scientific and technological innovation talents (Grant No. 2021RC4025), and the China Scholarship Council (202206130080).

References

[1] L. Song, H. Sun, J. Liu, Z. Yu, C. Cui, Automatic segmentation and quantification of global cracks in concrete structures based on deep learning, Measurement. 199 (2022) 111550, https://doi.org/10.1016/j.measurement.2022.111550.
[2] P.O. Pinheiro, R. Collobert, From Image-level to Pixel-level Labeling with Convolutional Networks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015: pp. 1713–1721. https://uhn.idm.oclc.org/login?url=http://ovidsp.ovid.com/ovidweb.cgi?T=JS&CSC=Y&NEWS=N&PAGE=fulltext&D=cctr&AN=CN-01587372 http://nt2yt7px7u.search.serialssolutions.com/?sid=OVID:Cochrane+Central+Register+of+Controlled+Trials&genre=article&id=pmid:844961.
[3] C.Z. Dong, F.N. Catbas, A review of computer vision–based structural health monitoring at local and global levels, Struct. Heal. Monit. 20 (2021) 692–743, https://doi.org/10.1177/1475921720935585.
[4] J.-M. Guo, H. Markoni, J.-D. Lee, BARNet: Boundary Aware refinement network for crack detection, IEEE Trans. Intell. Transp. Syst. 23 (7) (2022) 7343–7358, https://doi.org/10.1109/TITS.2021.3069135.
[5] Y.J. Cha, W. Choi, O. Büyüköztürk, Deep learning-based crack damage detection using convolutional neural networks, Comput. Civ. Infrastruct. Eng. 32 (2017) 361–378, https://doi.org/10.1111/mice.12263.
[6] L. Zhang, J. Shen, B. Zhu, A research on an improved Unet-based concrete crack detection algorithm, Struct. Heal. Monit. 20 (2021) 1864–1879, https://doi.org/10.1177/1475921720940068.
[7] W. Qiao, B. Ma, Q. Liu, X. Wu, G. Li, Computer vision-based bridge damage detection using deep convolutional networks with expectation Maximum attention module, Sensors. 21 (2021) 824, https://doi.org/10.3390/s21030824.
[8] Y. Tang, A.A. Zhang, L. Luo, G. Wang, E. Yang, Pixel-level pavement crack segmentation with encoder-decoder network, Measurement. 184 (2021) 109914, https://doi.org/10.1016/j.measurement.2021.109914.
[9] J. Pang, H. Zhang, H. Zhao, L. Li, DcsNet: A real-time deep network for crack segmentation, Signal, Image Video Process. 16 (2022) 911–919, https://doi.org/10.1007/s11760-021-02034-w.
[10] W. Jiang, M. Liu, Y. Peng, L. Wu, Y. Wang, HDCB-net: A neural network with the hybrid dilated convolution for pixel-level crack detection on concrete bridges, IEEE Trans. Ind. Informatics. 17 (2021) 5485–5494, https://doi.org/10.1109/TII.2020.3033170.

[11] Y. Yan, S. Zhu, S. Ma, Y. Guo, Z. Yu, CycleADC-Net: A crack segmentation method based on multi-scale feature fusion, Measurement. 204 (2022) 112107, https://doi.org/10.1016/j.measurement.2022.112107.

[12] T. Liu, L. Zhang, G. Zhou, W. Cai, C. Cai, L. Li, A.M. Mosa, BC-DUnet-based segmentation of fine cracks in bridges under a complex background, PLoS One. 17 (3) (2022), https://doi.org/10.1371/journal.pone.0265258.

[13] C. Xiang, W. Wang, L. Deng, P. Shi, X. Kong, Crack detection algorithm for concrete structures based on super-resolution reconstruction and segmentation network, Autom. Constr. 140 (2022) 104346, https://doi.org/10.1016/j.autcon.2022.104346.

[14] G. Li, Q. Liu, S. Zhao, W. Qiao, X. Ren, Automatic crack recognition for concrete bridges using a fully convolutional neural network and naive Bayes data fusion based on a visual detection system, Meas. Sci. Technol. 31 (7) (2020) 075403, https://doi.org/10.1088/1361-6501/ab79c8.

[15] H. Chen, H. Lin, M. Yao, Improving the efficiency of encoder-decoder architecture for pixel-level crack detection, IEEE Access. 7 (2019) 186657–186670, https://doi.org/10.1109/ACCESS.2019.2961375.

[16] X. Gao, C. Huang, S. Teng, G. Chen, A deep-convolutional-neural-network-based semi-supervised learning method for anomaly crack detection, Appl. Sci. 12 (2022) 9244, https://doi.org/10.3390/app12189244.

[17] J. Guo, Q. Wang, S. Su, Y. Li, Informativeness-guided active learning for deep learning–based façade defects detection, Comput. Civ. Infrastruct. Eng. (2023) 1–18, https://doi.org/10.1111/mice.12998.

[18] Z. Al-Huda, B. Peng, R.N.A. Algburi, S. Alfasly, T. Li, Weakly supervised pavement crack semantic segmentation based on multi-scale object localization and incremental annotation refinement, Appl. Intell. (2022) 1–20, https://doi.org/10.1007/s10489-022-04212-w.

[19] S. Shim, J. Kim, G.C. Cho, S.W. Lee, Multiscale and adversarial learning-based semi-supervised semantic segmentation approach for crack detection in concrete structures, IEEE Access. 8 (2020) 170939–170950, https://doi.org/10.1109/ACCESS.2020.3022786.

[20] H. Wang, Y. Li, L.M. Dang, S. Lee, H. Moon, Pixel-level tunnel crack segmentation using a weakly supervised annotation approach, Comput. Ind. 133 (2021) 103545, https://doi.org/10.1016/j.compind.2021.103545.

[21] X. Yang, R. Chen, F. Zhang, L. Zhang, X. Fan, Q. Ye, L. Fu, Pixel-level automatic annotation for forest fire image, Eng. Appl. Artif. Intell. 104 (2021) 104353, https://doi.org/10.1016/j.engappai.2021.104353.

[22] W. Wang, C. Su, Semi-supervised semantic segmentation network for surface crack detection, Autom. Constr. 128 (2021) 103786, https://doi.org/10.1016/j.autcon.2021.103786.

[23] Y. Shi, J. Yang, Z. Qi, Unsupervised anomaly segmentation via deep feature reconstruction, Neurocomputing. 424 (2021) 9–22, https://doi.org/10.1016/j.neucom.2020.11.018.

[24] S. Noor, M. Waqas, M.I. Saleem, H.N. Minhas, Automatic object tracking and segmentation using unsupervised siammask, IEEE Access. 9 (2021) 106550–106559, https://doi.org/10.1109/ACCESS.2021.3101054.

[25] G. Li, J. Wan, S. He, Q. Liu, B. Ma, Semi-supervised semantic segmentation using adversarial learning for pavement crack detection, IEEE Access. 8 (2020) 51446–51459, https://doi.org/10.1109/ACCESS.2020.2980086.

[26] J. Zhu, J. Song, Weakly supervised network based intelligent identification of cracks in asphalt concrete bridge deck, Alexandria Eng. J. 59 (2020) 1307–1317, https://doi.org/10.1016/j.aej.2020.02.027.

[27] Y. Zheng, M. Yang, M. Wang, X. Qian, R. Yang, X. Zhang, W. Dong, Semi-supervised adversarial semantic Segmentation network using transformer and multiscale convolution for High-resolution remote sensing imagery, Remote Sens. 14 (2022) 1–21, https://doi.org/10.3390/rs14081786.

[28] A. Tarvainen, H. Valpola, Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results, in: in: 31st Conf. Neural Inf. Process. Syst. (NIPS 2017), Curran Associates Inc, Long Beach, CA, USA, 2017, pp. 1–10.

[29] T. Vu, M. Cord, P. Patrick, ADVENT : Adversarial Entropy Minimization for Domain Adaptation in Semantic Segmentation, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019: pp. 2517–2526. https://openaccess.thecvf.com/content_CVPR_2019/supplemental/Vu_ADVENT_Adversarial_Entropy_CVPR_2019_supplemental.pdf.

[30] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D.P. Hughes, D.Z. Chen, Deep Adversarial Networks for Biomedical Image Segmentation Utilizing Unannotated Images, in: M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D.L. Collins, S. Duchesne (Eds.), Med. Image Comput. Comput. Assist. Interv. – MICCAI 2017, Springer, 2017, pp. 408–416, https://doi.org/10.1007/978-3-319-66179-7_47.

[31] L. Yu, S. Wang, X. Li, C.W. Fu, P.A. Heng, Uncertainty-Aware Self-ensembling Model for Semi-supervised 3D Left Atrium Segmentation, in: Int. Conf. Med. Image Comput. Comput. Interv., Springer, 2019: pp. 605–613. https://doi.org/10.1007/978-3-030-32245-8_67.

[32] V. Verma, K. Kawaguchi, A. Lamb, J. Kannala, A. Solin, Y. Bengio, D. Lopez-Paz, Interpolation consistency training for semi-supervised learning, Neural Networks. 145 (2022) 90–106, https://doi.org/10.1016/j.neunet.2021.10.008.

[33] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, S. Zhang, Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency, in, Int. Conf. Med. Image Comput. Comput. Interv., Springer (2021) 318–329, https://doi.org/10.1007/978-3-030-87196-3_30.

[34] X. Chen, Y. Yuan, G. Zeng, J. Wang, Semi-Supervised Semantic Segmentation with Cross Pseudo Supervision, in: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2021: pp. 2613–2622. https://doi.org/10.1109/cvpr46437.2021.00264.

[35] G. Chen, J. Ru, Y. Zhou, I. Rekik, Z. Pan, X. Liu, Y. Lin, B. Lu, J. Shi, MTANS: Multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation, Neuroimage. 244 (2021) 118568, https://doi.org/10.1016/j.neuroimage.2021.118568.

[36] J. Guo, Q. Wang, Y. Li, Semi-supervised learning based on convolutional neural network and uncertainty filter for façade defects classification, Comput. Civ. Infrastruct. Eng. 36 (2021) 302–317, https://doi.org/10.1111/mice.12632.

[37] D. He, K. Xu, Z. Peng, D. Zhou, Surface defect classification of steels with a new semi-supervised learning method, Opt. Lasers Eng. 117 (2019) 40–48, https://doi.org/10.1016/j.optlaseng.2019.01.011.

[38] E. Karaaslan, U. Bagci, F.N. Catbas, Attention-guided analysis of infrastructure damage with semi-supervised deep learning, Autom. Constr. 125 (2021) 103634, https://doi.org/10.1016/j.autcon.2021.103634.

[39] Y. Liu, J.K.W. Yeoh, Vision-Based Semi-Supervised Learning Method for Concrete Crack Detection, in: Constr. Res. Congr. 2020 Comput. Appl. - Sel. Pap. from Constr. Res. Congr. 2020, American Society of Civil Engineers Reston, VA, 2020: pp. 527–536. https://doi.org/10.1061/9780784482865.056.

[40] S. Shim, J. Kim, S.W. Lee, G.C. Cho, Road damage detection using super-resolution and semi-supervised learning with generative adversarial network, Autom. Constr. 135 (2022) 104139, https://doi.org/10.1016/j.autcon.2022.104139.

[41] G. Li, X. Li, J. Zhou, D. Liu, W. Ren, Pixel-level bridge crack detection using a deep fusion about recurrent residual convolution and context encoder network, Measurement. 176 (2021) 109171, https://doi.org/10.1016/j.measurement.2021.109171.

[42] Y. Xie, J. Zhang, C. Shen, Y. Xia, CoTr: Efficiently Bridging CNN and Transformer for 3D Medical Image Segmentation, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 12903 LNCS (2021) 171–180. https://doi.org/10.1007/978-3-030-87199-4_16.

[43] A. Lou, X. Yao, Z. Liu, J. Noble, Min-Max Similarity : A Contrastive Learning Based Semi-Supervised Learning Network for Surgical Tools Segmentation, Arxiv Print. (2022) arXiv:2203.15177. https://doi.org/10.48550/arXiv.2203.15177.

[44] X. Zhao, C. Fang, D. Fan, X. Lin, F. Gao, G. Li, Cross-Level Contrastive Learning and Consistency Constraint for Semi-Supervised Medical Image Segmentation, in: 19th Int. Symp. Biomed. Imaging, IEEE, 2022: pp. 1–5. https://doi.org/10.1109/ISBI52829.2022.9761710.

[45] X. Luo, M. Hu, T. Song, G. Wang, S. Zhang, Semi-Supervised Medical Image Segmentation via Cross Teaching between CNN and Transformer, ArXiv Prepr. (2021) arXiv:2112.04894. https://arxiv.org/abs/2112.04894.

[46] Y. Zhang, H. Liu, Q. Hu, TransFuse: Fusing Transformers and CNNs for Medical Image Segmentation, in: Med. Image Comput. Comput. Assist. Interv., Strasbourg, France, 2021: pp. 14–24. https://doi.org/10.1007/978-3-030-87193-2_2.

[47] E. Asadi Shamsabadi, C. Xu, D. Dias-da-Costa, Robust crack detection in masonry structures with Transformers, Measurement. 200 (2022) 111590, https://doi.org/10.1016/j.measurement.2022.111590.

[48] Wang, P. Cao, J. Wang, O.R. Zaiane, UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer, in: Proc. AAAI Conf. Artif. Intell., 2022: pp. 2441–2449. https://doi.org/10.1609/aaai.v36i3.20144.

[49] D.H. Kang, Y.J. Cha, Efficient attention-based deep encoder and decoder for automatic crack segmentation, Struct. Heal. Monit. 21 (2022) 2190–2205, https://doi.org/10.1177/14759217211053776.

[50] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Proc. 31st Int. Conf. Neural Inf. Process. Syst., Long Beach California, USA, 2017: pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762.

[51] D. Jha, M.A. Riegler, D. Johansen, P. Halvorsen, H.D. Johansen, DoubleU-Net: A deep convolutional neural network for medical image segmentation, in, Proc. - IEEE Symp. Comput. Med. Syst. (2020) 558–564, https://doi.org/10.1109/CBMS49503.2020.00111.

[52] H. Wu, S. Chen, G. Chen, W. Wang, B. Lei, Z. Wen, FAT-Net: Feature adaptive transformers for automated skin lesion segmentation, Med. Image Anal. 76 (2022) 102327, https://doi.org/10.1016/j.media.2021.102327.

[53] J.S. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Deep Residual Learning for Image Recognition, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Las Vegas Nevada, USA, 2016: pp. 770–778. https://doi.org/10.1002/chin.200650130.

[54] S. Feng, H. Zhao, F. Shi, X. Cheng, M. Wang, Y. Ma, D. Xiang, W. Zhu, X. Chen, Cpfnet: Context pyramid fusion network for medical image segmentation, IEEE Trans. Med. Imaging. 39 (2020) 3008–3018, https://doi.org/10.1109/TMI.2020.2983721.

[55] Z. Gu, J. Cheng, H. Fu, K. Zhou, H. Hao, Y. Zhao, T. Zhang, S. Gao, J. Liu, CE-Net: Context encoder network for 2D medical image segmentation, IEEE Trans. Med. Imaging. 38 (2019) 2281–2292, https://doi.org/10.1109/TMI.2019.2903562.

[56] C. Feng, H. Zhang, H. Wang, S. Wang, Y. Li, Automatic pixel-level crack detection on dam surface using deep convolutional network, Sensors (Switzerland). 20 (2020) 2069, https://doi.org/10.3390/s20072069.

[57] Z. Qu, J. Mei, L. Liu, D.Y. Zhou, Crack detection of concrete pavement with cross-entropy loss function and improved VGG16 network model, IEEE Access. 8 (2020) 54564–54573, https://doi.org/10.1109/ACCESS.2020.2981561.

[58] Y. Bao, J. Li, T. Nagayama, Y. Xu, B.F. Spencer, H. Li, A summary and benchmark problem, Struct. Heal. Monit. 20 (4) (2021) 2229–2239, https://doi.org/10.1177/14759217211006485.

[59] G. Li, Q. Liu, W. Ren, W. Qiao, B. Ma, J. Wan, Automatic recognition and analysis system of asphalt pavement cracks using interleaved low-rank group convolution hybrid deep network and SegNet fusing dense condition random field,

Measurement. 170 (2021) 108693, https://doi.org/10.1016/j.measurement.2020.108693.
[60] Y. Liu, J. Yao, X. Lu, R. Xie, L. Li, DeepCrack: A deep hierarchical feature learning architecture for crack segmentation, Neurocomputing. 338 (2019) 139–153, https://doi.org/10.1016/j.neucom.2019.01.036.
[61] Y. Gao, M. Zhou, D. Metaxas, UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation, in: Med. Image Comput. Comput. Assist. Interv. Springer International Publishing, Strasbourg, France (2021) 61–71, https://doi.org/10.1007/978-3-030-87199-4_6.