Smart Structures and Systems, *Vol. 33, No. 6 (2024) 449-463*
https://doi.org/10.12989/sss.2024.33.6.449

A label-free high precision automated crack detection method based on unsupervised generative attentional networks and swin-crackformer

Shiqiao Meng [a], Lezhi Gu [b], Ying Zhou* and Abouzar Jafari [c]

State Key Laboratory of Disaster Reduction in Civil Engineering, College of Civil Engineering, Tongji University, 1239 Siping Rd., Shanghai, 200092, China

(*Received May 23, 2023, Revised July 30, 2024, Accepted August 22, 2024*)

Abstract. Automated crack detection is crucial for structural health monitoring and post-earthquake rapid damage detection. However, realizing high precision automatic crack detection in the absence of corresponding manual labeling presents a formidable challenge. This paper presents a novel crack segmentation transfer learning method and a novel crack segmentation model called Swin-CrackFormer. The proposed method facilitates efficient crack image style transfer through a meticulously designed data preprocessing technique, followed by the utilization of a GAN model for image style transfer. Moreover, the proposed Swin-CrackFormer combines the advantages of Transformer and convolution operations to achieve effective local and global feature extraction. To verify the effectiveness of the proposed method, this study validates the proposed method on three unlabeled crack datasets and evaluates the Swin-CrackFormer model on the METU dataset. Experimental results demonstrate that the crack transfer learning method significantly improves the crack segmentation performance on unlabeled crack datasets. Moreover, the Swin-CrackFormer model achieved the best detection result on the METU dataset, surpassing existing crack segmentation models.

Keywords: crack detection; deep learning; unsupervised generative attentional networks; vision Transformer

1. Introduction

Crack detection plays an essential role in structural disaster prevention and mitigation, such as rapid post-earthquake damage assessment and structural health monitoring. The degree of cracks' development reflects the damage to the building. Traditionally, crack detection was carried out manually by observing the cracks' location, length, and width. The main drawbacks of traditional crack detection methods are low accuracy, poor comprehend-siveness, personnel safety risk, and especially low efficiency. It is necessary to seek ways to detect the crack morphology as accurately and quickly as possible, especially in post-earthquake disaster assessment and rescue efforts. Considering the apparent visual features of cracks, image processing was used to detect cracks to overcome the disadvantages mentioned above in manual detection. With the continuous development of image processing technology, the effect of crack detection has been gradually improved. Traditional image processing methods' effect in different environments is greatly affected by the parameters set manually, which leads to poor generalization capability and susceptibility to environmental interference. Researchers use traditional machine learning for crack detection, but limited by the fitting ability of the model, the detection accuracy in complex engineering environments is relatively low. The development of deep learning algorithms has dramatically improved the fitting ability of the model and has a stronger generalization ability. From the perspective of deep learning -algorithm, the crack detection task can be divided into three categories: crack image classification (Cha *et al*. 2017), crack object detection (Gou *et al*. 2019, Du *et al*. 2021), and crack semantic segmentation. The first two tasks are of low difficulty, and now they have achieved relatively high accuracy, even surpassing human recognition accuracy in classification problems (He *et al*. 2015). However, these algorithms can only roughly pinpoint the location of the crack instead of the crack geometric information, thus providing less information on disaster prevention and mitigation. Semantic crack segmentation is crack detection at the pixel level, which can detect crack geometric information such as shape, length, and width. The crack geometric information helps to reflect the type and degree of building damage. Rapid post-earthquake damage assessment often relies on detailed information, so the semantic segmentation of crack images is essential.

Existing crack segmentation algorithms fully utilized the advantages of convolutional neural networks (CNN) and proposed a variety of methods based on UNet (Ronneberger *et al*. 2015), DeepLab v3+ (Chen *et al*. 2018) and other classical models. Transformer architecture (Vaswani *et al*. 2017) achieved great success in Natural Language

*Corresponding author, Ph.D., Professor,
E-mail: yingzhou@tongji.edu.cn
[a] Ph.D. Student
[b] Undergraduate Student
[c] Ph.D.
[d] Ph.D.

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

Processing (NLP) and has recently been applied to computer vision tasks. Compared to CNN, Vision Transformer (ViT) (Dosovitskiy *et al*. 2020) relies on excellent modeling capabilities to achieve superior performance on multiple benchmarks such as ImageNet, COCO, and ADE20k. Recently, researchers proposed a few approaches based on Transformer methods like ViT (Ma *et al*. 2023) and Swin Transformer (Guo *et al*. 2023) in the field of crack segmentation.

However, the model based on Transformer focuses more on acquiring global information and less on local information than CNN (Raghu *et al*. 2021). Although the crack semantic segmentation achieves superb detection results with various algorithmic improvements utilizing supervised learning, obtaining pixel-level labeling annotations in the practical post-earthquake scenario is always time-consuming and challenging. The model is expected to have better migration ability, so the crack detection can be completed faster after the earthquake through the existing model and the unlabeled crack images collected. In the past, when dealing with similar tasks, a pre-trained model based on the existing annotated dataset (source domain) was directly applied to unannotated post-earthquake image data (target domain). When the difference in style between the source domain and target domain is significant, directly using the model trained in the source domain to detect cracks in the target domain will cause inferior crack detection results due to the limitation of the model's generalization ability. As a result, the effect of rapid post-earthquake damage detection is relatively insufficient.

In order to solve the above problems, this paper proposes a Swin-CrackFormer model that combines the advantages of CNN and Transformer. The proposed Swin-CrackFormer is based on UNet's cross-layer connection architecture and introduces the multi-head self-attention mechanism of the Swin-Transformer model into the encoder to achieve better global information perception. In addition, this paper proposes a local attention module that uses CNN to obtain more detailed local information perception, thereby obtaining higher crack detection accuracy. Furthermore, this paper proposes a crack segmentation transfer learning method, which implements domain transfer between two crack datasets via a generative adversarial network (GAN) to unify images of different styles and then utilizes annotations from existing datasets. Through this method, we implement the conversion between crack images in target domains and source domains, realizing high-precision migration of the existing crack segmentation algorithm among different domains. Through the data preprocessing method proposed in this article, the generation effect of GAN on local features of cracks can be effectively improved, thereby achieving higher-precision crack detection. It is worth mentioning that this is the first time that the GAN-based domain adaptation method has been applied to the practical post-earthquake crack detection scenario. To verify the effectiveness of the crack segmentation transfer learning method and the crack segmentation algorithm Swin-CrackFormer proposed in this paper, we utilized four crack datasets: METU, Crack500, DJITongji, and LD dataset. Among them, we set the METU dataset as the source domain and the others as the target domain for experiments, respectively. In addition, we conducted comparative experiments on these datasets and compared the Swin-CrackFormer proposed in this paper with existing high-precision crack detection algorithms through supervised training, thus verifying the effectiveness of the crack segmentation algorithm proposed in this paper.

In summary, the main contributions of this paper are as follows:

(1) This paper presents a novel framework for crack segmentation transfer learning. By employing preprocessing techniques for high-resolution data and leveraging GAN-based methods to generate synthetic datasets, the proposed framework offers an efficient solution for achieving high-precision crack detection in situations where labeled data is lacking. The effectiveness of this data processing method has been thoroughly validated, demonstrating significant improvements in the detection results.
(2) This paper proposes a high-precision crack segmentation algorithm, Swin-CrackFormer, which combines the architectural advantages of CNN and Transformer. Local information is efficiently acquired using the proposed local attention module, while efficient global information extraction is achieved through Swin Transformer blocks, resulting in crack segmentation with higher pixel-level accuracy.
(3) This paper presents the DJITongji and LD datasets collected from ordinary buildings and earthquake-damaged buildings. The crack images in these data sets are finely annotated at the pixel level, providing a reliable evaluation standard for daily structural health monitoring and rapid crack damage detection after disasters.

2. Related works

Some areas related to our research are introduced, and the connections between them and our research are clarified.

2.1 Crack semantic segmentation

Due to the obvious visual features of crack images, traditional image processing algorithms used in early research, segmented cracks with manually defined features. The processing methods include threshold segmentation and morphology (Subirats *et al*. 2006, Wang *et al*. 2007), various filtering techniques (Oliveira and Correia 2009, Salman *et al*. 2013), and edge detection algorithm (Zhao *et al*. 2010). Jahanshahi *et al*. (2013) conducted pioneering research in the field of image-based high-precision crack detection. They initially employed image-based three-dimensional scene reconstruction techniques. Subsequently, they utilized morphological processing and machine learning algorithms, including support vector machines (SVM), for the classification of cracks and quantification of crack widths. Sun *et al*. (2022) proposed a curvature filtering improved crack detection method via graph-based

anomalies calculation. However, the above traditional image processing algorithms need to define parameters manually, which is a complicated process. Their generalization performance is weak and easily disturbed by the environment. With the rapid development of machine learning methods, researchers began to take advantage of machine learning's powerful fitting capabilities and improved the crack detection performance over previous methods. Random structure forest (Shi *et al*. 2016) was used to segment the image patch, and the SVM or K-nearest neighbor (KNN) classifier was used to classify whether the patch contained cracks. Fernandes and Ciobanu (2014) used SVM to segment cracks based on previously extracted feature-based graphs. The main problem of the above machine learning-based methods is that their accuracy is limited when dealing with crack segmentation due to insufficient fitting ability.

With the development of computing power, convolutional neural networks (CNNs) based on deep learning have made remarkable achievements in crack detection. Makantasis *et al*. (2015) compared a multitude of traditional crack segmentation methods with a CNN architecture to segment tunnel wall defects. This research shows that using CNN to segment defect areas outperforms previous methods, such as SVM, KNN, and classification trees. Besides, with the proposal of various new CNN network structures, some structures and details suitable for crack segmentation have been created and added. Zhu *et al*. (2022) proposed an end-to-end crack image segmentation framework based on a one-step CNN for pixel-level object recognition with high accuracy on steel bridge cracks. Wang and Xiang (2021) proposed a two-stage method merging edge detection and CNN to reduce the burden of computing for detecting cracks in railway sleepers with high accuracy.

Many recent crack segmentation networks are based on encoders and decoders. Inside the encoder, the convolution and downsampling operations create denser, spatially smaller feature maps, which are then upsampled in the decoder to match the dimensions of the inputs to obtain segmentation results. Jenkins *et al*. (2018) used a fully convolutional network (FCN) based on a primitive encoder-decoder structure. The DeepCrack architecture proposed by Zou *et al*. (2018) is based on SegNet, which includes a multi-scale fusion component to extract and utilize features from multiple scales of the feature pyramid to generate segmentation maps. Another DeepCrack proposed by Liu *et al*. (2019) consists of the extended FCN and the Deeply-Supervised Nets (DSN). Meng *et al*. (2020) proposed Grid-DeepLab, which models the sub-regions of crack images with different importance features to endow the model with the ability to distinguish the effective area of the image.

With the development of the Transformer and its derivative Vision Transformer in computer vision, researchers began to use Transformer architecture on crack detection networks. Compared with CNN, Transformer focuses more on the correlation between basic visual elements and has higher visual semantic information. The Transformer-based method is used by Ma *et al*. (2023) to classify and detect pipeline defects, and the accuracy is improved compared with the CNN-based method. Guo *et al*. (2023) unifies Swin-Transformer as the encoder and the decoder with all multi-layer perception (MLP) layers for automatically detecting long and complicated pavement cracks. However, the CNN-based network has more significant advantages in extracting local features and visual structure than the Transformer. Therefore, combining the advantages of CNN and Transformer is a promising research direction to improve the accuracy of crack segmentation (Zhang *et al*. 2023).

The latest achievements of crack detection involve not only the integration and improvement of algorithms but engineering applications and setup of database and user interfaces. Meng *et al*. (2022b) proposed a three-stage crack detection method based on deep learning. The proposed three-stage method combines a preprocessing method, a classification method, a segmentation algorithm, and a postprocessing method to solve the problem that cracks only occupy a minimal area of the high-resolution image. Meng *et al*. (2022a) proposed an automated real-time crack detection method based on a drone, further improving crack detection accuracy in the engineering environment. Zou *et al*. (2022) used YOLOv4 to detect multicategory damage, including fine cracks, wide cracks, and concrete spalling. Moreover, a graphical user interface was developed to facilitate the post-earthquake reinforced concrete structural damage assessment.

However, although the methods mentioned above can realize high-precision crack detection under specific conditions, the reliable training effect depends on a large number of annotations. Applying the existing model directly to a new dataset or an actual post-earthquake scenario is ineffective.

2.2 Image transfer learning

Transfer learning aims to improve a learner's performance on particular objects in a target domain by transferring knowledge in a different but related source domain (Zhuang *et al*. 2020). Transfer learning has applications in various fields and is a promising field in machine learning, mainly dealing with images. Domain adaptation (DA) (Wang and Deng 2018) is a subclass of transfer learning, which maps data or features from different domains to the feature space of the target domain and uses them to enhance the training of the target domain to obtain better training results. Considering the absence of target domain labels in actual post-earthquake scenes and the similarity of crack morphology in different scenes, this task is an unsupervised DA problem, which belongs to homogeneous DA, with the following feasible methods. The method using deep learning starts with Maximum Mean Discrepancy (MMD) as statistic loss (Tzeng *et al*. 2014, Long *et al*. 2015, 2017). Later, an adversarial-based DA approach (Ganin and Lempitsky 2015) emerged, using adversarial loss. It draws on GAN's thought and uses H-divergence to construct a loss function. In addition, pseudo-labeling can be applied to achieve unsupervised domain adaptation (Saito *et al*. 2017, Chen *et al*. 2019). In terms of application, Siu *et al*. (2022) uses the non-adversarial deep learning network proposed by Johnson *et al*. (2016) for

synthetic data generation and augmentation to solve the data shortage problem of sewer defect detection. By adding a comparative learning module, defect detection accuracy is further improved. Reconstruction-based methods (Ghifary *et al*. 2016, Zhang *et al*. 2018) are similar to adversarial-based methods by constructing a reconstruction loss to reconstruct the source domain. The representative model of this type of method is mainly cycleGAN (Zhu *et al*. 2017), which has a good effect on transfer learning. Because there are result images of the intermediate reconstruction process, the reconstruction results can be evaluated intuitively.

2.3 Generative adversarial networks

GANs are powerful generative models that generate realistic-like samples without explicit real-world data distributions or mathematical conditions. These advantages allow GAN to be applied in various academic and engineering fields (Alqahtani *et al*. 2021, Liu *et al*. 2022). In the general architecture, GAN has two types of networks called discriminator and generator, denoted as D and G, respectively (Alqahtani *et al*. 2021). During the training phase, the weights of G are updated with a certain learning rate, while the weights of D are updated to reduce classification errors (Tyagi and Yadav 2021). The original GAN network only had the basic structure to generate and distinguish images from noise. Conditional GAN (Mirza and Osindero 2014) can make the results of GAN meet certain conditions, which can control the final output result by artificially changing the input vector. CycleGAN can transform images between different domains, providing a novel domain adaptation method based on GAN.

Various networks based on GAN have been applied to crack detection tasks, mainly for generating virtual datasets to augment datasets. Pei *et al*. (2021) combines the advantages of VAE and DCGAN and proposes an improved deep convolutional generative adversarial network (DCGAN) based virtual image dataset generation method for asphalt pavement cracks. In the research of Dunphy *et al*. (2022), GAN is used for data expansion to balance the data set, which can directly extract features for transfer learning, reduce the number of classified images input into the CNN network, and optimize the classification effect at the same time.

The improved GAN network represented by cycleGAN (Zhu *et al*. 2017) can realize the conversion between different styles of images, also known as domain adaptation. Chen *et al*. (2022) used cycleGAN to convert building information modeling (BIM) renderings into realistic images, and Hu *et al*. (2022) used it to make synthetic radar images more realistic. Using cycleGAN, Zhang *et al*. (2020) creates an unsupervised learning method for crack segmentation. Domain X (crack image) and Domain Y (Ground truth) are used as initial data, and Domain X and Y have no correspondence. After the training of cycleGAN, the crack labeling result is finally obtained. The segmentation effect is close to CNN, and the advantage of this unsupervised learning method is that it does not require manual labeling of crack segmentation datasets. Various applications have shown the effectiveness and practicability of GAN-based methods in domain adaptation. Therefore, this paper proposes a GAN-based method to convert the style of crack images, thereby significantly improving the migration ability to existing crack segmentation algorithms. In this way, the accuracy of crack detection can be greatly improved without manual labeling in actual engineering scenarios.

3. Methodology

3.1 Overall process

In this study, a GAN-based transfer learning method for crack detection and an improved crack semantic segmentation algorithm was developed. The core idea of the method is to fully use the data of the source domain and the target domain through the transfer learning method to achieve a significantly better crack segmentation effect without manually finely labeling the cracks in the target domain. It is crucial to emphasize that the main objective of this article is crack segmentation, distinct from the classification of crack images that can be accomplished through a conventional binary classification neural network (Chen and Jahanshahi 2017). The process of crack segmentation relies on pixel-level annotation, enabling precise quantification and localization of cracks. Consequently, the proposed crack transfer learning method leverages a dataset inherently comprised of crack instances as its default choice. Considering the self-similarity characteristics observed in cracks (Meng *et al*. 2022b), the extraction of sub-images from crack images does not result in significant information loss, further supporting the feasibility of this approach. Moreover, due to the inherent limitations of GANs in preserving local information during the generation of large scenes, this article adopts a strategic approach of dividing high-resolution images into multiple sub-images for subsequent operations. This approach aims to enhance the effectiveness of GAN-based synthesis in generating realistic synthetic images.

The overall process of crack detection involves the following steps: Firstly, it is necessary to acquire a source domain dataset that includes high-precision pixel-level annotations, which can be obtained from an open-source crack dataset. Additionally, obtaining an unannotated crack dataset specifically for the target domain images is also essential. Subsequently, the crack images undergo a gridding process, where high-resolution images are divided using a sliding window of size 512×512, resulting in the generation of numerous sub-images. These procedures collectively enable the creation of multiple original datasets, each comprising 512×512 images, which will be utilized for the transfer learning network.

Next, a pair of domains consisting of METU (source domain) paired with another dataset (target domain) is fed into the style transfer network. This paper uses the effective GAN model, called unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation (U-GAT-IT) (Kim *et al*. 2019), as a style transfer network. The synthetic image of the pair

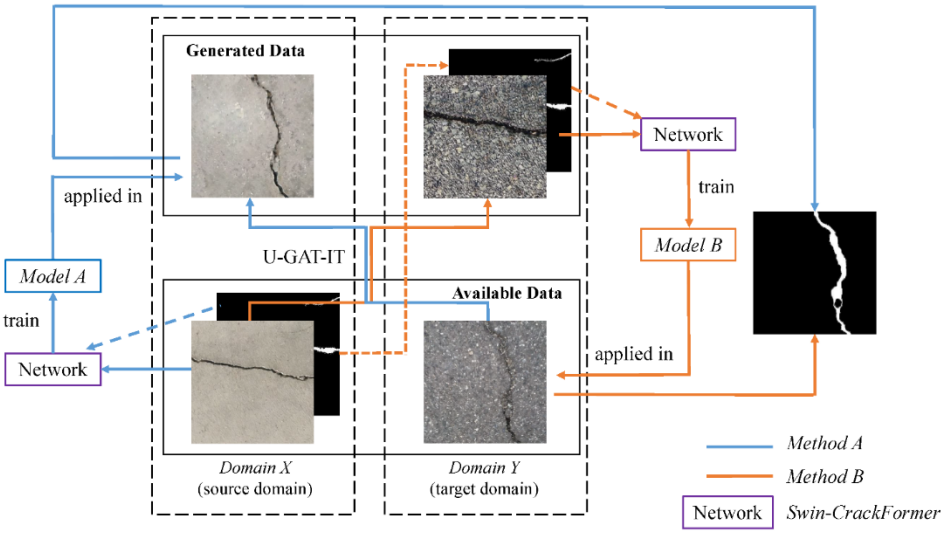


Fig. 1 Overview of the overall process. There are two methods of transfer learning based on U-GAT-IT. Method A: Applying the Model A which is trained by images and annotations in Domain X to fake images in Domain X; Model B: Applying the Model B which is trained by fake images in target domain and real annotations in source domain to real images in Domain Y. Both methods finally get the segmentation results of the target domain

of domains can be obtained through transfer learning, and the corresponding annotation of the image in the source domain can be used to annotate the synthetic dataset.

Finally, the labelling of the target domain is obtained using an enhanced crack segmentation algorithm. The sub-images are subsequently reassembled to restore the original image size. Two distinct methods, denoted as A and B in Fig. 1, are employed in this process. In practical engineering applications, method B necessitates the training of the segmentation network after acquiring the target domain image, whereas method A does not require model training during the testing phase. Method B, due to its additional model training, entails a relatively longer duration. By employing the aforementioned approach, significantly more accurate crack segmentation results can be achieved compared to utilizing only source domain data, eliminating the need for crack data labeling in the target domain.

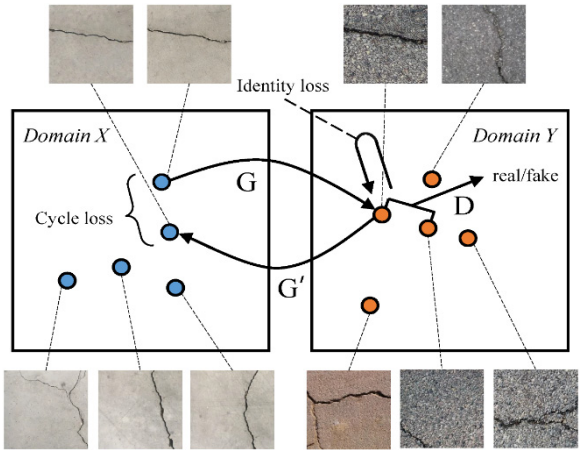


Fig 2 Overview of the approach using U-GAT-IT for crack image domain adaption. The figure only shows the transfer process from domain *X* to domain *Y*. *G*, *G*' are generators, and *D* represents the discriminator

3.2 U-GAT-IT

Fig. 2 introduces the core parts of our proposed method for domain adaptation using U-GAT-IT. The full name of U-GAT-IT is unsupervised generative attentional networks with adaptive layer instance normalization for image-to-image translation. U-GAT-IT aims to train a function G that maps images from a source domain X to a target domain Y using only unpaired samples drawn from each domain. We chose it as the network for domain adaption for its better performance in the comparison experiment on improving the segmentation effect.

The framework of U-GAT-IT consists of two generators $G_{X\to Y}$ and $G_{Y\to X}$ and two discriminators D_X and D_Y. The two generators generate synthetic images of source and target domain and the two discriminators are classifiers to detect the artificial images generated from the real images from corresponding domain. During the training phase, the generators' weights are updated at a certain learning rate, whereas the discriminators' weights are updated to decrease the classification error. To focus on the global critical information of the image better, U-GAT-IT integrates the attention module into both generator and discriminator. It should be noted that the specific structure of the network can be replaced by any generator or discriminator network structure. The specific structure used in this article comes from the research of Kim *et al*. (2019).

U-GAT-IT includes four losses, which are Adversarial loss, Cycle loss, Identity loss, and CAM loss respectively. The following parts describe their purposes and expressions.

Adversarial loss: This loss function is used to match the distribution of translated images to the target domain distribution, and the adversarial loss is the core of GAN-

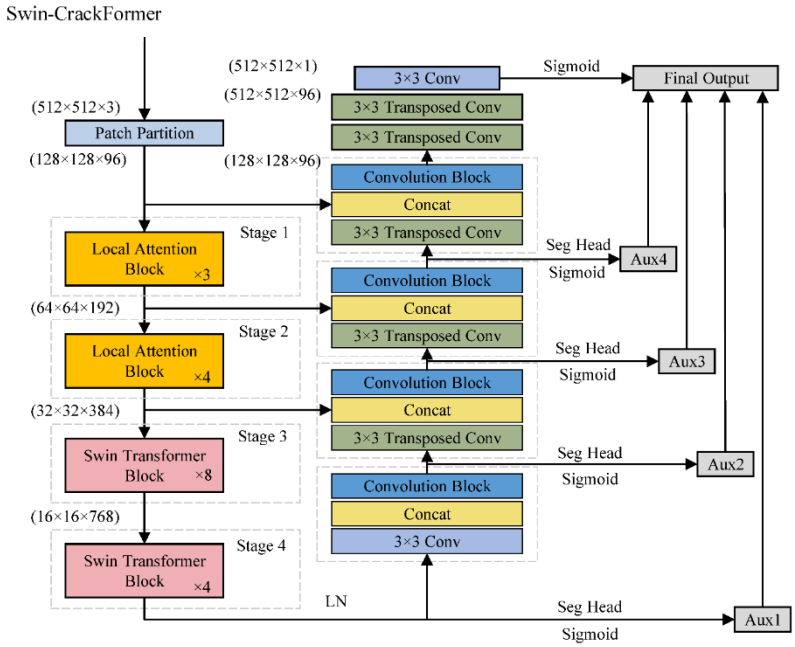


Fig. 3 The complete architecture of proposed Swin-CrackFormer

based networks. Optimizing this loss function helps make the generated images more similar to those in the target domain.

$$L_{gan}^{X\to Y} = \mathbb{E}_{x\sim Y}[D_Y(x)^2] + \mathbb{E}_{x\sim Y}[1 - D_Y(G_{X\to Y}(x))^2] \quad (1)$$

Cycle loss: Cycle loss was applied earlier by cycleGAN (Zhu *et al*. 2017), where it was called cycle-consistent loss. As the name suggests, this loss function ensures that the generated fracture structure pattern is consistent with the input fracture, thus alleviating the pattern collapse problem. In the absence of cycle loss, the generator is likely to learn a mapping between two domains that do not depend on the input domain X image while satisfying the requirements of the adversarial loss, which leads to the enormous change of the location and shape of cracks in the generated image. Thus, given an image $x \in X$, after the sequential translation of x from Domain X to Domain Y and from Domain Y to Domain X, the image should be successfully translated back to the original image domain. The expression of the cycle loss is shown in Eq. (2).

$$L_{cycle}^{X\to Y} = \mathbb{E}_{x\sim X}[|x - G_{Y\to X}(G_{X\to Y}(x))|_1] \quad (2)$$

Identity loss: To ensure that the color distributions of the input image and output image are similar, we apply an identity consistency constraint to the generator, whose expression is shown in Eq. (3). Given an image $x \in X$, after the translation of x using $G_{X\to Y}$ the image should not change.

$$L_{identity}^{X\to Y} = \mathbb{E}_{x\sim Y}[|x - G_{X\to Y}(x)|_1] \quad (3)$$

CAM loss: CAM Loss is a unique loss function of the U-GAT-IT model based on the complex network designed in the model. Given an image $x \in \{X, Y\}$, using the information given by the auxiliary classifiers η_x and η_{DY} in the model, $G_{X\to Y}$ and D_Y can know where they need to improve or what is the most significant difference between the two domains in their current state. This loss function can be expressed as follows

$$L_{cam}^{X\to Y} = -\big(\mathbb{E}_{x\sim X}\big[log(\eta_X(x))\big] + \mathbb{E}_{x\sim Y}[log(1 - \eta_X(x))]\big) \quad (4)$$

$$L_{cam}^{D_Y} = \mathbb{E}_{x\sim Y}[\eta_{D_Y}(x)^2] + \mathbb{E}_{x\sim X}[(1 - \eta_{D_Y}G_{X\to Y}(x))^2]) \quad (5)$$

Full objective: In the training phase of the model, the encoder, decoder, discriminator, and auxiliary classifier of U-GAT-IT are trained based on the above four loss functions, and the final optimization goal is as follows

$$L_{total} = \min_{G_{X\to Y},G_{Y\to X},\eta_X,\eta_Y} \max_{D_X,D_Y,\eta_{D_X},\eta_{D_Y}} (\lambda_1 L_{gan} + \lambda_2 L_{cycle} + \lambda_3 L_{identity} + \lambda_4 L_{cam}) \quad (6)$$

For the hyperparameters, we set $\lambda_1 = 1$, $\lambda_2 = 3$, $\lambda_3 = 3$, $\lambda_4 = 1000$. Here, $L_{gan} = L_{gan}^{X\to Y} + L_{gan}^{Y\to X}$ and the other losses are defined similarly. (L_{cycle}, $L_{identity}$ and L_{cam}).

3.3 Swin-CrackFormer

The Swin-CrackFormer proposed in this paper adopts a cross-layer connection architecture similar to the UNet network (Ronneberger *et al*. 2015), which can fully integrate high-level and low-level semantic information. On this basis, the respective advantages of CNN and Transformer are combined to obtain good results in extracting local and global features. The schematic diagram of the network structure of Swin-CrackFormer is shown in Fig. 3. First, it is necessary to use a convolution kernel with

a size of 4×4 and a stride of 4 for processing, thereby reducing the size of the feature map to reduce the amount of calculation. After that, four stages are needed to extract features. The first two stages focus on extracting local features, which are mainly stacked by local attention blocks. The latter two stages mainly aim to extract global features, which are implemented by stacking Swin Transformer blocks. After the first three stages, it is necessary to use a convolutional layer for downsample processing and simultaneously increase the number of channels. The channels used to calculate the four stages are 96, 192, 384, and 768, respectively.

3.3.1 Local attention block

The first two stages of Swin-CrackFormer proposed in this paper are stacked using local attention blocks. The first stage includes three local attention blocks, and the second includes four. The local attention block includes dynamic position encoding (DPE), multi-head local relationship aggregator (MHLRA), and feed-forward neural network (FFN). The network structure diagram is shown in Fig. 4(a). Among them, the dynamic position coding is mainly completed through depthwise separable convolution (DWConv), and the FFN is mainly composed of two convolutional layers, ReLU and two dropout layers. This paper designs MHLRA as a multi-head style, which means that each head processes a group of channel information independently. The channels of each group first generate a context token $V(X)$ through a convolutional layer and then aggregate the context under the influence of the local relationship parameter L. The value of the local relationship parameter L is only related to the relative position. The specific expression of MHLRA is shown in Eq. (7).

$$O_n(X) = L_n V_n(X) \tag{7}$$

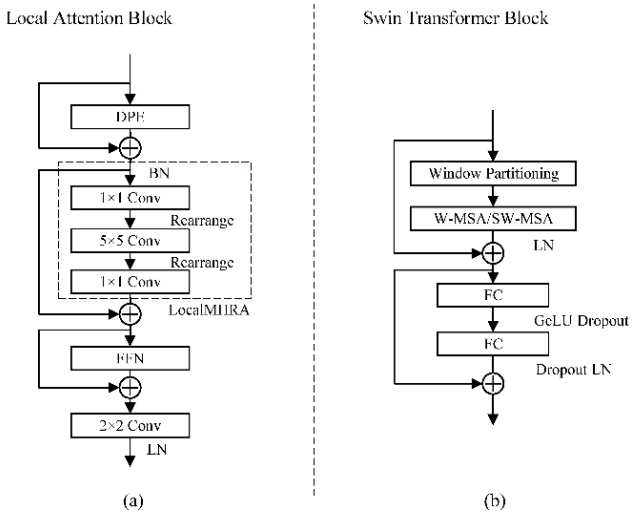


Fig. 4 The structure of Local Attention Block (a) and Swin-Transformer Block (b), which are used several times in the Swin-CrackFormer. LN: layer normalization; BN: batch normalization; DPE: dynamic position embedding; FC: fully connected layer; FFN: feedforward neural network

where X is the input feature map, $V(X)$ represents the context token after convolution, O represents the output feature maps, L_n indicates the local relationship parameter in one of the heads. The specific expression of L_n is shown in Eq. (8).

$$L_n(X_i, X_j) = l_n^{i-j} \tag{8}$$

where X_i is the anchor token, X_j is any token in the local neighborhood, l_n is the learnable parameter matrix, i-j is the relative position of the two tokens. The above operations can be implemented with convolutional layers. In each head, a parameter-learnable convolution layer is used as the specific value of the local relation parameter L, and the feature maps in each channel are processed equally. In this paper, the attention range set by the local attention block is 5 pixels in the neighborhood, which means that the convolution kernel size in the convolutional layer is 5×5. Finally, combining the results of all heads and using a convolutional layer for processing is necessary. The specific expression is shown in Eq. (9).

$$MHLRA(X) = Concat(O_1(X); O_2(X); \ldots; O_n(X))U \tag{9}$$

where *Concat* means splicing all heads' output in the channel dimension, U means mapping through a convolutional layer. In this paper, the number of heads used by MHLRA in the first two stages of Swin-CrackFormer is 1 and 3, respectively.

3.3.2 Swin TransFormer block

The last two stages of Swin-CrackFormer proposed in this paper are stacked using Swin TransFormer blocks, including 8 and 4 Swin TransFormer blocks, respectively. Before the start of each stage, a convolutional layer is used to downsample and increase the number of channels. The Swin TransFormer block mainly includes a window attention layer or a shifted window attention layer, and its network structure diagram is shown in Fig. 4(b). It should be noted that the number of blocks contained in a stage must be an even number. The reason is that the two modules must appear in pairs, which means a block containing a window attention module and a block containing a shifted window attention module need to appear alternately. The traditional Transformer directly calculates the global attention, so the computational complexity is exceptionally high. However, Swin Transformer limits the calculation of attention to each window, thereby reducing the amount of calculation. In the window attention module, the feature map needs to be divided first, and the size of each window is 16×16. When calculating window attention, it is necessary to convert the two-dimensional feature map in each window into a one-dimensional feature vector, which converts the 16×16 feature map into a vector with a length of 256 through reshaping. After that, the feature map is split from the channel dimension and converted into multiple attention heads. Then, three fully connected layers map the feature vectors into key, query, and value vectors. Since these pixels have a specific positional relationship when performing attention calculations in each window, the

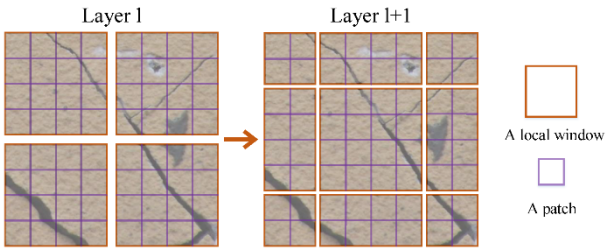


Fig. 5 An illustration of the shifted window approach for computing self-attention, which is calculated for each divided window. In layer l, windows are divided in a regular way, while the window partitioning is shifted in layer l+1. When calculating self-attention in the new window, the information of different windows in layer l is integrated, which enhances the fitting ability of the model

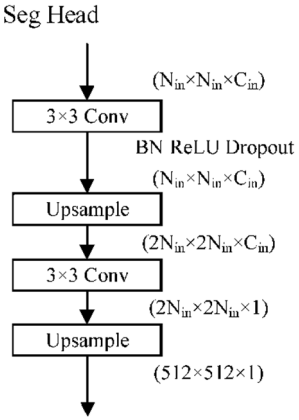


Fig. 6 The structure of Seg Head module. LN: layer normalization; BN: batch normalization; N_{in}: the dimension of the feature maps; C_{in}: the number of channels in the feature maps

positional relationship must be considered, so the relative position encoding B (Liu *et al*. 2021) is introduced. The calculation equation of window attention is shown in Eq. (10).

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d}} + B\right)V \qquad (10)$$

where Q, K, and V represent the query, key, and value vectors, respectively, d represents the number of channels, and B is the relative position code, whose value is only related to the relative position of the pixel and is a non-learnable parameter.

In order to better realize the information interaction between windows when calculating attention, Swin Transformer also introduces the shifted window operation. The schematic diagram illustrating the concepts is presented in Fig. 5. The left side of Fig. 5 depicts the window attention mechanism without overlapping, while the right side showcases the shift window attention mechanism, which involves the shifting of the window. It can be seen that the shifted window contains the elements of the original adjacent window. Therefore, in such manner, the global attention mechanism can be calculated with a relatively low amount of calculation to extract global information effectively. In this paper, the number of heads in the Swin Transformer block in the third and fourth stages is set to 12 and 24, respectively.

3.3.3 Other details

After the four-stage processing is completed, the feature map needs to be upsampled through the deconvolution layer to gradually restore the feature map to the size of the original image. After each upsample, it is necessary to use the cross-layer connection to fuse with the feature map output by the previous stage through concatenation. Then use the convolution block for mapping. The convolution block includes convolution layers, ReLU and batch normalization layers, and dropout layers. After merging with the feature maps of all previous stages, two deconvolutions are used for upsampling to return the feature maps to their original size. Finally, reduce the number of channels to 1 through a convolutional layer, and use the Sigmoid function for processing to complete the output of the final result.

In order to further improve the effect of feature extraction, four Seg Head modules are added to Swin-CrackFormer, using feature maps of four resolutions as input, and the output result is a prediction map of the same size as the final prediction result. In the model training process, the final output of the model and the output of the four Seg Head modules are calculated separately. Using the ground truth to calculate the loss of several output results, adding them up, and then using backpropagation to optimize the parameters. The detailed structure of the Seg Head module is shown in Fig. 6. In the stage of model inference, there is no need to use the Seg head module, but only the final result needs to be output directly.

4. Experiments

4.1 Dataset description

To verify the effectiveness of the crack detection transfer learning method and the effect of the proposed Swin-CrackFormer, we use four crack datasets: METU, Crack500, DJITongji, and LD datasets. The images in the METU dataset were captured in different buildings located at Middle East Technical University (Özgenel 2019). The whole dataset includes 458 images of concrete surface cracks. The size of each image is 4032×3024 pixels. Crack500 dataset (Zhang *et al*. 2016) is a public pavement crack dataset of 3264×2448 pixels collected by a low-cost smartphone.

The DJITongji dataset is collected by aerial photography over the buildings of Tongji University using a drone and is first presented in this paper. The size of the cracks is relatively small, and the wall styles in the crack pictures are relatively diverse. In addition, the crack images are accompanied by more interference factors. The size of the original image is 3840×2160, and there are 1902 images in total. This data set includes two tasks, crack classification and crack segmentation, which can provide a good

Table 1 The basic information of the datasets

Datasets	Number of images		Size of images	
	Original	Data-augmented	Original	Data-augmented
METU	458	3450	4032×3024	512×512
DJITongji	1902	5758	3840×2160	
Crack500	250	1514	3264×2448	
LD	419	1307	5184×3888	

evaluation benchmark for daily structural health monitoring.

The LD dataset is a dataset of cracks in earthquake-damaged buildings collected on-site after the Luding earthquake in Sichuan, China. Thus, it can be used as a benchmark for rapid damage detection after an earthquake. The size of the original image is 5184×3888, and there are 419 high-quality annotated images. Visible cracks can be seen in various patterns, including differences in color, wall material, and degree of spalling. In the above datasets, we set the METU dataset as the source domain and the other datasets as the target domains for experiments. In the comparative experiment for the Swin CrackFormer model, we use the METU dataset for comparative experiments.

The images in these datasets are eventually scaled to 512×512 pixels to match the situation applied in the actual detection. A series of steps were carried out to process the obtained original data sets: (1) Crop the images. The size of the sub-images obtained was 1024×1024, and the overlapped part of the crop process was 512 pixels in both horizontal and vertical directions. (2) Select appropriate images by setting the upper and lower bounds of the number of crack pixels. Images with few crack pixels are difficult to use in the style transfer learning network and even cause tremendous interference to the model's training. Images with too numerous crack pixels often correspond to huge peeling, so adding them will also interfere with the model. (3) Resize the chosen images to 512×512 and rename these images.

4.2 Evaluation metrics

The evaluation metrics used in this experiment are mean recall, mean precision, and mean intersection over union (mIoU). The expressions of the three evaluation standards are shown in Eq. (11)-(13).

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$mIoU = \frac{TP}{TP + FP + FN} \tag{13}$$

where *TP* and *FP* represent the total number of pixels that actually label to cracks or not cracks when the prediction results are cracks; *FN* represents the total number of pixels that the prediction result are non-cracks, but the actual annotation are cracks.

4.3 Implementation details

In order to save video random access memory (VRAM) and calculation time, we chose the light version of U-GAT-IT to train and test in this experiment. For the experiment of the crack segmentation transfer learning method, the following hyperparameter values were used: learning rate = 0.0001, batch size = 1, weight decay factor = 0.0001. For each pair of domains, a total of 100000 iterations have been learned. The model and loss function are introduced in Section 3.2.2. For the comparison experiment of Swin-CrackFormer, the learning rate during the training process of all models is 1×10^{-5}, and the optimizer used in this experiment is Adam (Kingma and Ba 2014). The number of training epochs is 500. The loss function used during model training is Dice loss added to binary cross entropy loss. The graphics processing unit (GPU) used in model training and inference in all experiments is Tesla V100. All code writing is completed using the PyTorch framework (Paszke *et al*. 2019).

4.4 Experiments of the crack segmentation transfer learning method

In this experiment, U-GAT-IT is first used to perform crack image transfer learning through two methods to obtain converted images. Then combine the generated images with the annotations of the original dataset to make a new dataset and use Swin-CrackFormer for crack segmentation. In order to demonstrate the effect of the transfer learning method, this paper demonstrates the gradual change of the U-GAT-IT model generation effect during the training process, shown in Fig. 7. We also visually compare the segmentation effects of three target domain crack images under supervised learning, two transfer learning methods and directly test methods in Fig. 8. The "directly test" method involves training the model exclusively on the source domain and subsequently testing it directly on the target domain. By employing this method, we can compare the detection results obtained and evaluate the effectiveness of the proposed transfer learning approach.

Table 2 shows the performance of the transfer learning method for crack segmentation on the Crack500 dataset. This dataset is an asphalt crack dataset, which is characterized by fairly similar image styles throughout the dataset, large crack widths, and little environmental disturbance. The results show that the *Recall* of directly test method is high, but the *Precision* is extremely low, indicating that there are too many erroneous points identifying non-cracks as cracks in the prediction results of directly test method. Method A has a significant improvement in *Precision*, which dramatically improves *mIoU*. Method B performs better than method A in *mIoU*. However, because method B needs to train the Swin-CrackFormer model, the time consumption of the overall process takes longer. The *mIoU* of method B is close to supervised learning under completely unlabeled and unsupervised conditions, which is more than 34% higher than the directly test method.

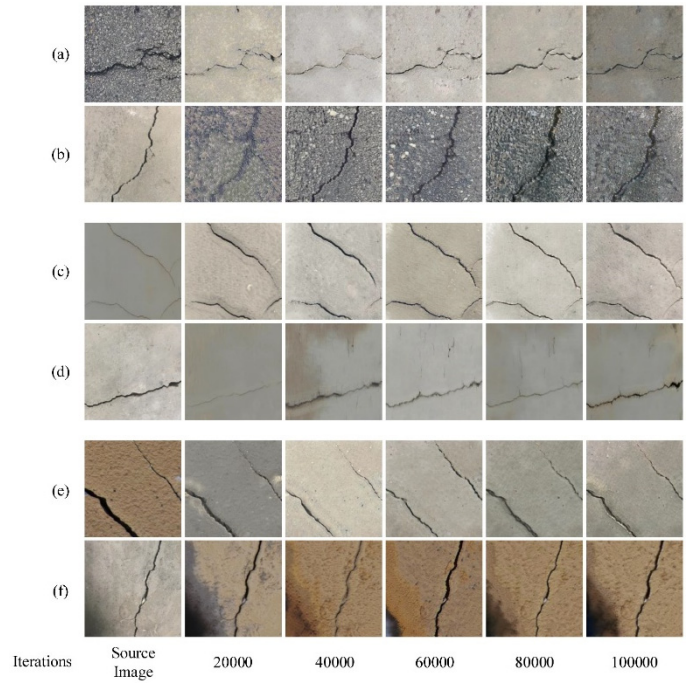


Fig. 7 Comparison of the conversion results of different pairs of domains on different iterations: (a) From Crack500 to METU dataset; (b) From METU to Crack500 dataset; (c) From DJITongji to METU dataset; (d) From METU to DJITongji dataset; (e) From LD to METU dataset; (f) From METU to LD dataset

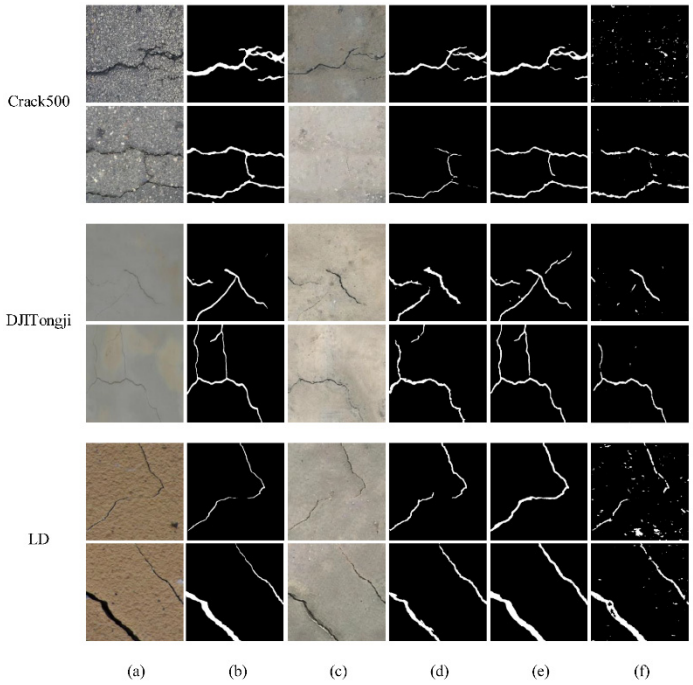


Fig. 8 Comparison of segmentation effects of supervised learning, two transfer learning methods and directly test method. (a) Images of the three target domains; (b) Segmentation results of supervised training using Swin-CrackFormer; (c) Source domain images obtained by style transfer using U-GAT-IT; (d) The segmentation results of Method A; (e) The segmentation results of Method B, (f) The segmentation results of the directly test method

Table 3 shows the performance of the crack segmentation transfer learning method on the DJITongji dataset. The cracks in this dataset are usually thin and have some environmental interference. Compared with the directly test method, the proposed method achieves better results on several evaluation metrics. The *mIoU* increases by more than 10% for both methods. Compared with Method A, Method B has better *Precision* and slightly lower *Recall*, resulting in higher *mIoU*. It can be seen from the experimental results that using the method proposed in

Table 2 Experiment of the crack segmentation transfer learning method on the Crack500 dataset

Evaluation metric	Supervised learning	Method A	Method B	Directly test
Recall (%)	82.42	46.34	57.01	84.84
Precision (%)	76.74	79.29	80.96	18.07
mIoU (%)	65.69	40.54	48.76	14.64

Table 3 Experiment of the crack segmentation transfer learning method on the DJITongji dataset

Evaluation metric	Supervised learning	Method A	Method B	Directly test
Recall (%)	68.92	46.37	38.97	26.17
Precision (%)	75.73	44.95	55.82	44.73
mIoU (%)	55.79	27.91	29.10	17.40

Table 4 Experiment of the crack segmentation transfer learning method on the LD dataset

Evaluation metric	Supervised learning	Method A	Method B	Directly test
Recall (%)	74.73	56.73	65.84	86.18
Precision (%)	85.90	61.26	56.51	13.57
mIoU (%)	67.02	40.60	44.25	12.10

this paper can greatly improve the crack detection accuracy of daily structural health detection in unknown environments.

Table 4 shows the performance of the proposed transfer learning method for crack segmentation on the LD dataset. The images in this dataset have different crack widths, styles, and background theme colors. The results show that, similar to the Crack500 dataset, the directly test method has higher *Recall* but lower *Precision*. Both methods significantly improve accuracy. In such a post-earthquake scenario, compared with the directly test method, the *mIoU* of the two methods is improved by more than 28%, and method B improves by 32.15%, which is more than 3.5 times the *mIoU* of the directly test method. The experiment results greatly verify the effectiveness of our proposed method in practical post-earthquake crack detection scenarios.

4.5 Comparative experiments of the crack segmentation model

In this experiment, this paper optimizes the model parameters through supervised learning on the METU dataset and compares the model's effect when cracks are labeled. The effect comparison models we chose include PSPNet (Zhao *et al*. 2017), Attention UNet (Oktay *et al*. 2018), DeepLab v3+ (Chen *et al*. 2018) and Swin-UNet (Cao *et al*. 2023), all of which are recent representative high-precision segmentation models.

PSPNet uses ResNet50 as the backbone. In PSPNet, the original image is downsampled into a feature map, input to the Pyramid Pooling Module (PPM) module, and then summed with the output. Finally, the result is obtained by convolution and bilinear interpolation upsampling. In addition, an auxiliary loss is added to the network to further improve the segmentation performance. Attention UNet is improved based on UNet, adding the Attention Gate module and implementing the attention mechanism for the skip connection and upsampling layers. The attention mechanism in this model is used to suppress irrelevant information in the image and highlight important local features. DeepLab v3+ uses an encoder-decoder architecture, and the backbone of the model uses ResNet152. The Swin-UNet builds a U-shaped symmetric encoder-decoder architecture with skip connections based on the Swin Transformer block. The model is built entirely based on the self-attention mechanism in Transformer.

Table 5 Comparative experiment of the Swin-CrackFormer on the METU dataset

Model	Recall (%)	Precision (%)	mIoU (%)	Parameters (M)
PSPNet	77.57	85.10	68.15	53.32
Attention UNet	80.98	86.37	71.74	59.19
DeepLab v3+	81.15	87.29	72.43	59.34
Swin-UNet	79.61	88.29	71.87	21.36
Swin-CrackFormer	84.89	89.28	76.99	61.69

Compared with the above four classical segmentation models, our proposed Swin-CrackFormer is trained on the METU dataset, and the evaluation metrics of the segmentation results are compared in Table 5. The experimental findings validate the efficacy of Swin-CrackFormer, as it consistently outperforms other models on all three evaluation metrics. These results provide empirical evidence of the superior performance achieved by the Swin-CrackFormer architecture. Additionally, it is noteworthy that the parameter count of the Swin-CrackFormer model is comparable to that of conventional models, including PSPNet, Attention UNet, and DeepLab v3+. The specific parameter quantities of these models are reported in Table 5, denoted in millions (M). Despite the Swin-CrackFormer model's more intricate architecture, its inference speed remains largely unaffected in practical applications. This implies that the Swin-CrackFormer model achieves a favorable balance between model complexity and computational efficiency.

4.6 Experiments of the impact of data preprocessing methods on the accuracy of crack transfer learning methods

The primary objective of this experiment is to validate the crack detection accuracy of the proposed crack transfer learning method on distinct preprocessed datasets. As discussed in Section 3.1, this article employs a technique to partition a high-resolution large image into multiple sub-images, aiming to enhance the effectiveness of the GAN in generating local features with higher precision. To

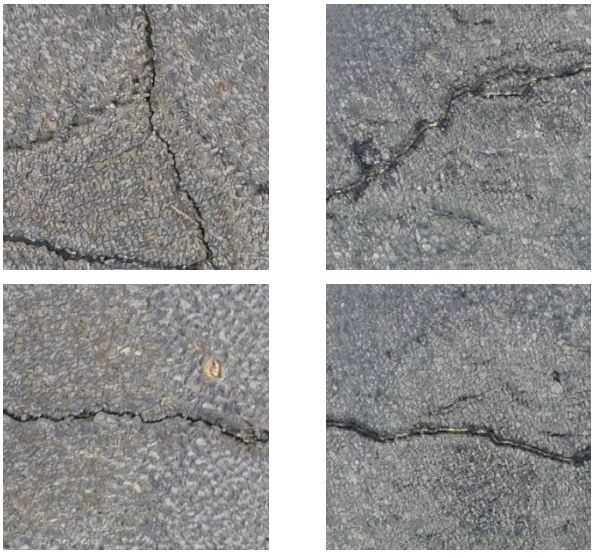
substantiate this hypothesis, an ablation experiment was conducted to investigate the impact of the transfer learning process.

In this experiment, the U-GAT-IT model is employed for the GAN framework, while the Swin-CrackFormer model is utilized for crack segmentation. However, two distinct data processing methods are employed. The first method involves dividing the large image into sub-images, as previously described. The second method, on the other hand, employs the resized original image directly. It is important to note that the primary distinction between these two methods lies in the level of detail captured in the images. The first method, referred to as the "close-up image dataset", offers greater detail, whereas the second method, known as the "distant image dataset", provides a broader global perspective. To facilitate clear understanding, Fig. 9 provides a visual representation of the synthesized images generated by the GAN using the aforementioned data from the two mentioned datasets, thereby presenting a comprehensive illustration of the visualization results. To ensure consistency in computational requirements and video memory usage, both image types are uniformly resized to 512×512 dimensions. Furthermore, to maintain fairness in the comparison, an equal number of images is selected from both datasets, and the evaluation is performed on the same test set.

The two datasets were trained and tested using the Method A and Method B of the proposed crack transfer learning framework. The original source dataset used in this experiment is the METU dataset and the original target dataset is the Crack500 dataset. The corresponding experimental results are presented in Table 6. The results demonstrate that, in the case of the close-up image dataset, both Method A and Method B exhibit significantly superior transfer learning performance compared to the distant image dataset. This provides substantial evidence supporting the efficacy of the data processing method proposed in this article for crack transfer learning. Moreover, these findings serve as validation for the rationality and effectiveness of the transfer learning framework presented in this study.

4.7 Ablation experiments of the structure of the Swin-CrackFormer model

In this experiment, we conducted an ablation study to investigate the structure of the proposed Swin-CrackFormer model, which incorporates two key modules: the local attention block and the Swin Transformer block. These modules play a crucial role in facilitating effective local and global feature extraction. Therefore, it is imperative to verify the efficacy of these blocks. To assess their impact, we replaced the local attention blocks and Swin Transformer blocks with convolutional layers and evaluated their influence on crack segmentation performance. Notably, the convolutional layer employed in our study included essential modules such as the batch norm layer and downsample layer, ensuring consistency with the modules utilized in UNet (Ronneberger *et al*. 2015). During each model training iteration, we only replaced one of the two aforementioned modules. The experimental results, presented in Table 7, unequivocally demonstrate the notable positive influence of both the local attention block and Swin Transformer block on the performance of the Swin-CrackFormer model, which decisively confirm the effectiveness of the local attention blocks and Swin Transformer blocks employed in this study.



(a) (b)

Fig. 9 Visualization of synthetic data produced by the close-up image dataset and the distant image dataset via transfer learning methods. (a) the generate result using the close-up image dataset; (b) the generate result using the distant image dataset

Table 6 Experiments of the impact of data preprocessing methods on the accuracy of crack transfer learning methods

Generated dataset	Transfer learning method	Recall (%)	Precision (%)	mIoU (%)
Distant image dataset	A	42.92	73.92	35.80
	B	37.61	79.96	33.34
Close-up image dataset	A	46.34	79.29	40.54
	B	57.01	80.96	48.76

Table 7 Ablation experiment of the structure of the Swin-CrackFormer on the METU dataset

Model	Recall (%)	Precision (%)	mIoU (%)
Swin-CrackFormer (without local attention blocks)	84.30	87.21	74.91
Swin-CrackFormer (without Swin Transformer blocks)	83.07	88.81	75.11
Swin-CrackFormer	84.89	89.28	76.99

5. Conclusions

This paper proposes a crack segmentation transfer learning method and a segmentation network that integrate vision transformer and convolution operation called Swin-CrackFormer. Using the method proposed in this paper, cracks can be identified more accurately and effectively without manual annotations. For the case of refined crack segmentation, higher precision of crack segmentation can be achieved. The proposed method has carried out a large number of experiments on multiple public datasets and the crack detection dataset proposed in this paper. Based on the experiment results, the conclusion of this paper is summarized as follows:

(1) This paper presents a novel crack segmentation transfer learning method that utilizes an efficient data preprocessing approach and leverages GAN models to generate high-quality synthetic data. The method narrows the domain gap between different datasets relative to the direct transfer of models on different datasets through domain adaptation so that the training and appliance of the segmentation models take place on the unified style crack images. The effectiveness of the data preprocessing method used was fully verified through ablation study. The method significantly improves the effectiveness and accuracy of crack segmentation in actual engineering scenarios.

(2) This paper proposed the Swin-CrackFormer, which combines the advantages of convolutional neural network and Transformer to balance local and global feature extraction better, thereby achieving higher crack segmentation accuracy. In the Swin-CrackFormer, we assimilate various structures and techniques. The feature extraction downsampling process is realized in four stages containing local attention blocks and Swin Transformer blocks to extract local and global information. Cross-layer connected network structure and Seg Head modules as auxiliary paths are applied to the aggregate extracted information. Through these designs, the Swin-CrackFormer can achieve an excellent segmentation effect.

(3) We conducted experiments in four datasets to verify our crack segmentation transfer learning method, selecting the METU dataset as the source domain and Crack500, DJITongji, and LD datasets as the target domain, respectively. The experimental results show that compared with the directly test method, Method A, which does not require segmentation training after transfer learning, increases the mIoU by more than 10.5% in DJITongji, more than 25.9% in Crack500, and more than 28.5% in LD. As for Method B, which requires a segmentation training process after transfer learning, further improves the above segmentation effects compared with Method A, and the experimental results fully reflect the advantages of our proposed method.

(4) This paper compares the proposed Swin-CrackFormer with various classical crack segmentation models under supervised learning conditions through comparative experiments. The experimental results reveal that Swin-CrackFormer outperforms many recent advanced segmentation models on *Precision*, *Recall*, and *mIoU* metrics. Through this comparative experiment, we verified that Swin-CrackFormer could achieve the best effect in the case of supervised learning. In addition, as an optimized crack segmentation method, Swin-CrackFormer can replace other segmentation networks in the complete transfer learning process, and its effectiveness has been demonstrated by experiments on multiple datasets, including LD datasets after an actual earthquake.

Acknowledgments

The research described in this paper was financially supported by the Distinguished Young Scientist Fund of National Natural Science Foundation of China (Grant No. 52025083), the Shanghai Social Development Science and Technology Research Project (Grant No. 22dz1201400), and the National Natural Science Foundation of China (Grant No. U2139209)

References

Alqahtani, H., Kavakli-Thorne, M. and Kumar, G. (2021), "Applications of generative adversarial networks (gans): An updated review", *Arch. Computat. Methods Eng.*, **28**, 525-552. https://doi.org/10.1007/s11831-019-09388-y

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. and Wang, M. (2023), "Swin-unet: Unet-like pure transformer for medical image segmentation", *Proceedings of Computer Vision–ECCV 2022 Workshops* (*Part III*), Tel Aviv, Israel, October.

Cha, Y.J., Choi, W. and Büyüköztürk, O. (2017), "Deep learning-based crack damage detection using convolutional neural networks", *Comput.-Aided Civil Infrastr. Eng.*, **32**(5), 361-378. https://doi.org/10.1111/mice.12263

Chen, F.C. and Jahanshahi, M.R. (2017), "NB-CNN: Deep learning-based crack detection using convolutional neural network and Nave Bayes data fusion", *IEEE Transact. Indust. Electr.*, **65**(5), 4392-4400. https://doi.org/10.1109/TIE.2017.2764844

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*).

Chen, C., Xie, W., Huang, W., Rong, Y., Ding, X., Huang, Y., Xu, T. and Huang, J. (2019), "Progressive feature alignment for unsupervised domain adaptation", *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chen, G., Teng, S., Lin, M., Yang, X. and Sun, X. (2022), "Crack detection based on generative adversarial networks and deep learning", *KSCE J. Civil Eng.*, **26**(4), 1803-1816. https://doi.org/10.1007/s12205-022-0518-2

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

Du, Y., Pan, N., Xu, Z., Deng, F., Shen, Y. and Kang, H. (2021), "Pavement distress detection and classification based on YOLO network", *Int. J. Pave. Eng.*, **22**(13), 1659-1672. https://doi.org/10.1080/10298436.2020.1714047

Dunphy, K., Sadhu, A. and Wang, J. (2022), "Multiclass damage detection in concrete structures using a transfer learning-based generative adversarial networks", *Struct. Control Health Monitor.*, **29**(11), e3079. https://doi.org/10.1002/stc.3079

Fernandes, K. and Ciobanu, L. (2014), "Pavement pathologies classification using graph-based features", In: *2014 IEEE International Conference on Image Processing* (*ICIP*).

Ganin, Y. and Lempitsky, V. (2015), "Unsupervised domain adaptation by backpropagation", *International Conference on Machine Learning*.

Ghifary, M., Kleijn, W.B., Zhang, M., Balduzzi, D. and Li, W. (2016), "Deep reconstruction-classification networks for unsupervised domain adaptation", *Proceedings of the 14th European Conference of Computer Vision-ECCV 2016* (*Part IV 14*), Amsterdam, The Netherlands, October.

Gou, C., Peng, B., Li, T. and Gao, Z. (2019), "Pavement crack detection based on the improved faster-rcnn", In: *2019 IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering* (*ISKE*).

Guo, F., Qian, Y., Liu, J. and Yu, H. (2023), "Pavement crack detection based on transformer network", *Autom. Const.*, **145**, 104646. https://doi.org/10.1016/j.autcon.2022.104646

He, K., Zhang, X., Ren, S. and Sun, J. (2015), "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification", *Proceedings of the IEEE International Conference on Computer Vision*.

Hu, D., Chen, J. and Li, S. (2022), "Reconstructing unseen spaces in collapsed structures for search and rescue via deep learning based radargram inversion", *Autom. Const.*, **140**, 104380. https://doi.org/10.1016/j.autcon.2022.104380

Huang, X. and Belongie, S. (2017), "Arbitrary style transfer in real-time with adaptive instance normalization", *Proceedings of the IEEE International Conference on Computer Vision*.

Jahanshahi, M.R., Masri, S.F., Padgett, C.W. and Sukhatme, G.S. (2013), "An innovative methodology for detection and quantification of cracks through incorporation of depth perception", *Mach. Vis. Appl.*, **24**, 227-241. https://doi.org/10.1007/s00138-011-0394-0

Jenkins, M.D., Carr, T.A., Iglesias, M.I., Buggy, T. and Morison, G. (2018), "A deep convolutional neural network for semantic pixel-wise segmentation of road and pavement surface cracks", In: *2018 26th European Signal Processing Conference* (*EUSIPCO*).

Johnson, J., Alahi, A. and Fei-Fei, L. (2016), "Perceptual losses for real-time style transfer and super-resolution", *Proceedings of the 14th European Conference of Computer Vision–ECCV 2016* (*Part II 14*), Amsterdam, The Netherlands, October.

Kim, J., Kim, M., Kang, H. and Lee, K. (2019), "U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation", arXiv preprint arXiv:1907.10830.

Kingma, D.P. and Ba, J. (2014), "Adam: A method for stochastic optimization", arXiv preprint arXiv:1412.6980.

Liu, Y., Yao, J., Lu, X., Xie, R. and Li, L. (2019), "DeepCrack: A deep hierarchical feature learning architecture for crack segmentation", *Neurocomputing*, **338**, 139-153. https://doi.org/10.1016/j.neucom.2019.01.036

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. and Guo, B. (2021), "Swin transformer: Hierarchical vision transformer using shifted windows", *Proceedings of the IEEE/CVF International Conference on Computer Vision*.

Liu, G., Niu, Y., Zhao, W., Duan, Y. and Shu, J. (2022), "Data anomaly detection for structural health monitoring using a combination network of GANomaly and CNN", *Smart Struct. Syst., Int. J.*, **29**(1), 53-62. https://doi.org/10.12989/sss.2022.29.1.053

Long, M., Cao, Y., Wang, J. and Jordan, M. (2015), "Learning transferable features with deep adaptation networks", *International Conference on Machine Learning*.

Long, M.S., Zhu, H., Wang, J.M. and Jordan, M.I. (2017), "Deep Transfer Learning with Joint Adaptation Networks", In: *International Conference on Machine Learning*, 2208-2217.

Ma, D., Fang, H., Wang, N., Lu, H., Matthews, J. and Zhang, C. (2023), "Transformer-optimized generation, detection, and tracking network for images with drainage pipeline defects", *Comput.-Aided Civil Infrastruct. Eng.*, **38**(15), 2109-2127. https://doi.org/10.1111/mice.12970

Makantasis, K., Protopapadakis, E., Doulamis, A., Doulamis, N. and Loupos, C. (2015), "Deep convolutional neural networks for efficient vision-based tunnel inspection", In: *2015 IEEE International Conference on Intelligent Computer Communication and Processing* (*ICCP*).

Meng, S., Zhang, X., Qiao, S. and Zhou, Y. (2020), "Research on grid optimized crack detection model based on deep learning", *J. Build. Struct.*, **41**(S2), 404-410.

Meng, S., Gao, Z., Zhou, Y., He, B. and Djerrad, A. (2022a), "Real-time automatic crack detection method based on drone", *Comput.-Aided Civil Infrastruct. Eng.*, **38**(7), 849-872. https://doi.org/10.1111/mice.12918

Meng, S., Gao, Z., Zhou, Y., He, B. and Kong, Q. (2022b), "A three-stage deep-learning-based method for crack detection of high-resolution steel box girder image", *Smart Struct. Syst., Int. J.*, **29**(1), 29-39. https://doi.org/10.12989/sss.2022.29.1.029

Mirza, M. and Osindero, S. (2014), "Conditional generative adversarial nets", arXiv preprint arXiv:1411.1784.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y. and Kainz, B. (2018), "Attention u-net: Learning where to look for the pancreas", arXiv preprint arXiv:1804.03999.

Oliveira, H. and Correia, P.L. (2009), "Automatic road crack segmentation using entropy and image dynamic thresholding", In: *2009 17th European Signal Processing Conference*.

Özgenel, C.F. (2019), "Concrete crack segmentation dataset", *Mendeley Data*, **1**, p. 2019.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N. and Antiga, L. (2019), "Pytorch: An imperative style, high-performance deep learning library", *Adv. Neural Inform. Process. Syst.*, **32**.

Pei, L., Sun, Z., Xiao, L., Li, W., Sun, J. and Zhang, H. (2021), "Virtual generation of pavement crack images based on improved deep convolutional generative adversarial network", *Eng. Applicat. Artif. Intell.*, **104**, 104376. https://doi.org/10.1016/j.engappai.2021.104376

Raghu, M., Unterthiner, T., Kornblith, S., Zhang, C. and Dosovitskiy, A. (2021), "Do vision transformers see like convolutional neural networks?", *Adv. Neural Inform. Process. Syst.*, **34**, 12116-12128.

Ronneberger, O., Fischer, P. and Brox, T. (2015), "U-net: Convolutional networks for biomedical image segmentation", *Proceedings of 18th International Conference on Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015* (*Part III 18*), Munich, Germany, October.

Saito, K., Ushiku, Y. and Harada, T. (2017), "Asymmetric tri-training for unsupervised domain adaptation", In: *International Conference on Machine Learning*.

Salman, M., Mathavan, S., Kamal, K. and Rahman, M. (2013), "Pavement crack detection using the Gabor filter", In: *The 16th International IEEE Conference on Intelligent Transportation Systems* (*ITSC 2013*).

Shi, Y., Cui, L., Qi, Z., Meng, F. and Chen, Z. (2016), "Automatic

road crack detection using random structured forests", *IEEE Transact*. *Intell*. *Transport*. *Syst*., **17**(12), 3434-3445. https://doi.org/10.1109/TITS.2016.2552248

Siu, C., Wang, M. and Cheng, J.C. (2022), "A framework for synthetic image generation and augmentation for improving automatic sewer pipe defect detection", *Autom*. *Const*., **137**, 104213. https://doi.org/10.1016/j.autcon.2022.104213

Subirats, P., Dumoulin, J., Legeay, V. and Barba, D. (2006), "Automation of pavement surface crack detection using the continuous wavelet transform", In: *2006 International Conference on Image Processing*.

Sun, W., Zhou, Y., Xiang, J., Chen, B. and Feng, W. (2022), "Crack detection in concrete slabs by graph-based anomalies calculation", *Smart Struct*. *Syst*., *Int*. *J*., **29**(3), 421-431. https://doi.org/10.12989/sss.2022.29.3.421

Tyagi, S. and Yadav, D. (2021), "A comprehensive review on image synthesis with adversarial networks: Theory, literature, and applications", *Arch*. *Computat*. *Methods Eng*., **29**, 2685-2705. https://doi.org/10.1007/s11831-021-09672-w

Tzeng, E., Hoffman, J., Zhang, N., Saenko, K. and Darrell, T. (2014), "Deep domain confusion: Maximizing for domain invariance", arXiv preprint arXiv:1412.3474.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017), "Attention is all you need", *Adv*. *Neural Inform*. *Process*. *Syst*., 30.

Wang, M. and Deng, W. (2018), "Deep visual domain adaptation: A survey", *Neurocomputing*, **312**, 135-153. https://doi.org/10.1016/j.neucom.2018.05.083

Wang, G. and Xiang, J. (2021), "Railway sleeper crack recognition based on edge detection and CNN", *Smart Struct*. *Syst*., *Int*. *J*., **28**(6), 779-789. https://doi.org/10.12989/sss.2021.28.6.779

Wang, K.C., Li, Q. and Gong, W. (2007), "Wavelet-based pavement distress image edge detection with a trous algorithm", *Transport*. *Res*. *Record*, **2024**(1), 73-81. https://doi.org/10.3141/2024-09

Zhang, L., Yang, F., Zhang, Y.D. and Zhu, Y.J. (2016), "Road crack detection using deep convolutional neural network", In: *2016 IEEE International Conference on Image Processing* (*ICIP*).

Zhang, W., Ouyang, W., Li, W. and Xu, D. (2018), "Collaborative and adversarial network for unsupervised domain adaptation", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zhang, K., Zhang, Y. and Cheng, H. (2020), "Self-supervised structure learning for crack detection based on cycle-consistent generative adversarial networks", *J*. *Comput*. *Civil Eng*., **34**(3), 04020004. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000883

Zhang, E., Shao, L. and Wang, Y. (2023), "Unifying transformer and convolution for dam crack detection", *Autom*. *Const*., **147**, 104712. https://doi.org/10.1016/j.autcon.2022.104712

Zhao, H., Qin, G. and Wang, X. (2010), "Improvement of canny algorithm based on pavement edge detection", In: *2010 3rd International Congress on Image and Signal Processing*.

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*.

Zhu, J.-Y., Park, T., Isola, P. and Efros, A.A. (2017), "Unpaired image-to-image translation using cycle-consistent adversarial networks", *Proceedings of the IEEE International Conference on Computer Vision*.

Zhu, H., Li, Z., Huang, M., Ji, P. and Zhang, Q. (2022), "One-step deep learning-based method for pixel-level detection of fine cracks in steel girder images", *Smart Struct*. *Syst*., *Int*. *J*., **29**(1), 153-166. https://doi.org/10.12989/sss.2022.29.1.153

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H. and He, Q. (2020), "A comprehensive survey on transfer learning", *Proceedings of the IEEE*, **109**(1), 43-76. https://doi.org/10.1109/JPROC.2020.3004555

Zou, Q., Zhang, Z., Li, Q., Qi, X., Wang, Q. and Wang, S. (2018), "Deepcrack: Learning hierarchical convolutional features for crack detection", *IEEE Transact*. *Image Process*., **28**(3), 1498-1512. https://doi.org/10.1109/TIP.2018.2878966

Zou, D., Zhang, M., Bai, Z., Liu, T., Zhou, A., Wang, X., Cui, W. and Zhang, S. (2022), "Multicategory damage detection and safety assessment of post-earthquake reinforced concrete structures using deep learning", *Comput*.*-Aided Civil Infrastruct*. *Eng*., **37**(9), 1188-1204. https://doi.org/10.1111/mice.12815