












Research Article

Hybrid Pixel-Level Crack Segmentation for Ballastless Track Slab Using Digital Twin Model and Weakly Supervised Style Transfer

Wenbo Hu ^{1,2}, Weidong Wang ^{3,4}, Xianhua Liu ^{3,4}, Jun Peng ^{3,4}, Sicheng Wang ^{3,4},
Chengbo Ai ⁵, Shi Qiu ^{3,4}, Wenjuan Wang ⁶, Jin Wang ^{3,4}, Qasim Zaheer ^{3,4},
and Lichang Wang ⁷

¹Department of Civil and Environmental Engineering, The Hong Kong Polytechnic University, Hong Kong 999077, China

²National Rail Transit Electrification and Automation Engineering Technology Research Center (Hong Kong Branch), Hong Kong 999077, China

³School of Civil Engineering, Central South University, Changsha 410075, China

⁴Center for Railway Infrastructure Smart Monitoring and Management, Central South University, Changsha 410075, China

⁵Department of Civil and Environmental Engineering, University of Massachusetts, Amherst, MA 01003, USA

⁶School of Business Administration, Capital University of Economics and Business, Beijing 100026, China

⁷School of Geosciences and Info-Physics, Central South University, Changsha 410075, China

Correspondence should be addressed to Shi Qiu; sheldon.qiu@csu.edu.cn

Received 10 July 2023; Revised 6 December 2023; Accepted 13 January 2024; Published 25 January 2024

Academic Editor: Sara Casciati

Copyright © 2024 Wenbo Hu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Crack detection at the pixel level across complex scenarios (structural interference and adverse working conditions) is a critical consideration in the maintenance of ballastless track slab (BTS). Although existing deep learning models achieve acceptable accuracy on cracks with a monotonous background, the ground truth with high labor cost is inevitable and their performance in complex scenarios may fall far below their theoretical bounds or even cause “all-black images.” A hybrid algorithm based on synthetic data from digital twin model and weakly supervised style transfer is proposed in this paper for addressing the above challenges. The algorithm uses a region-attention strategy to enable the uncontrolled generative adversarial network (GAN) focusing its attention on weak labels containing crack regions, directly obtaining segmentation results with the same style as the ground truth of the crack forest dataset. In addition, a digital twin model that can simulate the real inspection working conditions is established to generate a synthetic crack dataset, enabling the hybrid algorithm to extract the most discriminative features. The results show that the performance of the hybrid algorithm on inspection images across complex scenarios is nearly 25% higher than that of the DeepLabv3+ network, while the time cost consumed is only 0.5% of the latter. The deployment of the region attention strategy also enables the hybrid algorithm to achieve a mean intersection of union (MIoU) of 79.38%, which is nearly twice as much as that of GAN. It not only eliminates the oversegmentation caused by structures such as rails and fastener systems but also overcomes “all-black images.” In addition, synthetic data can greatly enlarge the range, type, and number of discriminative crack features compared with data augmentation based on limited real data, thus enhancing the performance of the hybrid algorithm for uncertain inspection data. Particularly, the fully trained hybrid algorithm based on the synthetic dataset shows good adaptability and generalization to adverse working conditions such as uneven lighting, noise, and blur.

1. Introduction

Ballastless track slab (BTS) of high-speed railway (HSR) is the important concrete structure that bears the dynamic load of rail, which deteriorates with service time increases [1, 2].

Distresses such as cracks not only reduce the strength of the track structure and shorten the service life of the BTS but may also cause the fastener falling off and the rail shifting, which threatens the operation safety of HSR [3–5]. Conventionally, experienced engineers conduct periodic visual

inspections to detect surface cracks of infrastructures and propose maintenance and rehabilitation strategies [6, 7]. However, manual inspection whose reliability depends on the experience of engineers and can only be performed at limited time windows (e.g., midnight), which is susceptible to missed or incorrect inspections due to complex environmental conditions (poor light, ambient interference, etc.). Therefore, there is an urgent need to achieve accurate and efficient pixel-level detection for cracks in BTS across complex backgrounds, which is also a key foundation for maintenance decisions.

Conventional machine vision-based solutions for automated pixel level crack detection have been proposed to replace the manual visual inspection, which obtain appreciable crack detection accuracy with only low computational cost by extracting and analyzing shallow or perceptible image features (color, grayscale, shape, edges, entropy, texture, histogram of oriented gradients, scale-invariant feature transform, etc.) [8–13]. Tang et al. [14] used fuzzy set theory and boundary histogram to determine the optimal threshold for distinguishing crack pixels and background pixels by maximizing the fuzzy index entropy. Oliveira and Correia [15] proposed a regional growth strategy capable of segmenting pavement cracks with complex shapes. The seed pixels obtained from the smoothed image through an effective segmentation procedure can minimize the prediction results of false positives. Xu et al. [16] first divided the binary image of the crack into subimages and extracted the parameters representing the crack characteristics from each subimage and then manually selected subimages with representative parameters to train the artificial neural network. Oliveira and Correia [17] characterized cracks based on image processing and pattern identification techniques to train KNN classifiers. However, noise and uneven lighting can adversely affect the computation of shallow features such as color and grayscale, resulting in blurred or discontinuous crack boundaries. In addition, the effectiveness of optimal thresholds and seed pixels relies on manual intervention, which is susceptible to less or oversegmentation results in detecting cracks with complex topology and low-contrast backgrounds [18].

Deep learning solutions represented by convolutional neural networks (CNNs) eliminate manual intervention in feature processing, which revolutionize the accuracy boundaries of traditional pixel-level crack detection solutions based on machine vision [19–22]. The CNNs rely on the convolutional layer (a large number of convolution kernels) inside the networks to perform a convolution operation with a neighborhood of the input crack image, which slides from the upper left to the lower right of the image with a certain step and outputs the deep abstract feature map for crack detection [23, 24]. Dorafshan et al. [25] compared the performance of DCNN-based pixel segmentation network with six commonly used edge detection methods (Roberts, Pre-witt, Sobel, Laplacian of Gaussian, Butterworth, and Gaussian) for the detection of concrete cracks. The results show that these edge detection methods based on heuristic

feature extraction can only detect 53–79% of the crack edge pixels, but DCNN can detect about 86% of the crack images through automatic feature extraction and has the fastest processing time. Liu et al. [26] used the U-Net full convolution network to identify the shape and location information of concrete cracks for the first time, which can achieve higher accuracy when using a smaller data set compared with DCNN and FCN-based methods. Zhang et al. [27] established a five-layer CNN-based pixel segmentation network (CrackNet) to detect cracks on 3D asphalt pavement and compared with 3D shadow modeling and SVM method based on HOG features in detection accuracy, which demonstrated the superiority of data-driven deep learning-based methods for crack detection at the pixel level. In addition, the limitations of convolutional filters in contextual information extraction from images prone to generate rough or discontinuous crack boundaries, especially for thin cracks. Ding et al. [28] proposed a visual transformer model with global self-attention mechanism, named IBR-Former, where boundary pixel information was applied to refine the segmentation, which improved the boundary location accuracy of thin cracks with complex shapes.

Although various pixel-level deep learning models have achieved extremely high theoretical accuracy on generic crack datasets (concrete structures such as building facades, pavements, bridges, track slabs, etc.) [29–32], three challenges posed by inspection images of BTS across complex scenarios cause these models to be deployed with accuracy far below their theoretical limits.

- (1) The complex scenarios of actual inspection images include structural interferences that are highly similar to cracks and adverse working conditions, as shown in Figure 1. Structures such as rails, fastener systems, and precracks can significantly disturb the identification of cracks resulting in oversegmentation compared with the monotonous concrete background of generic crack datasets. Moreover, adverse working conditions such as uneven lighting, noise, and blur can greatly reduce the differentiation of cracks from the background, which causes failure to extract the most discriminative features [33]. In addition, Kang et al. [34] also pointed out that there is almost no optimal pixel-level model that can finely segment cracks from such complex scenarios.
- (2) The extremely low proportion of crack pixels to the whole image, or even single-pixel skeleton, is also a key characteristic of inspection images across complex scenarios. These extremely low percentages of cracks are more likely to be misclassified as background, resulting in “all black images,” compared with the cracks across the whole image in generic crack datasets. Overfitting caused by the extreme imbalance of pixel proportions is also a significant factor for the poor accuracy of pixel-level deep learning models on inspection images across complex scenarios.

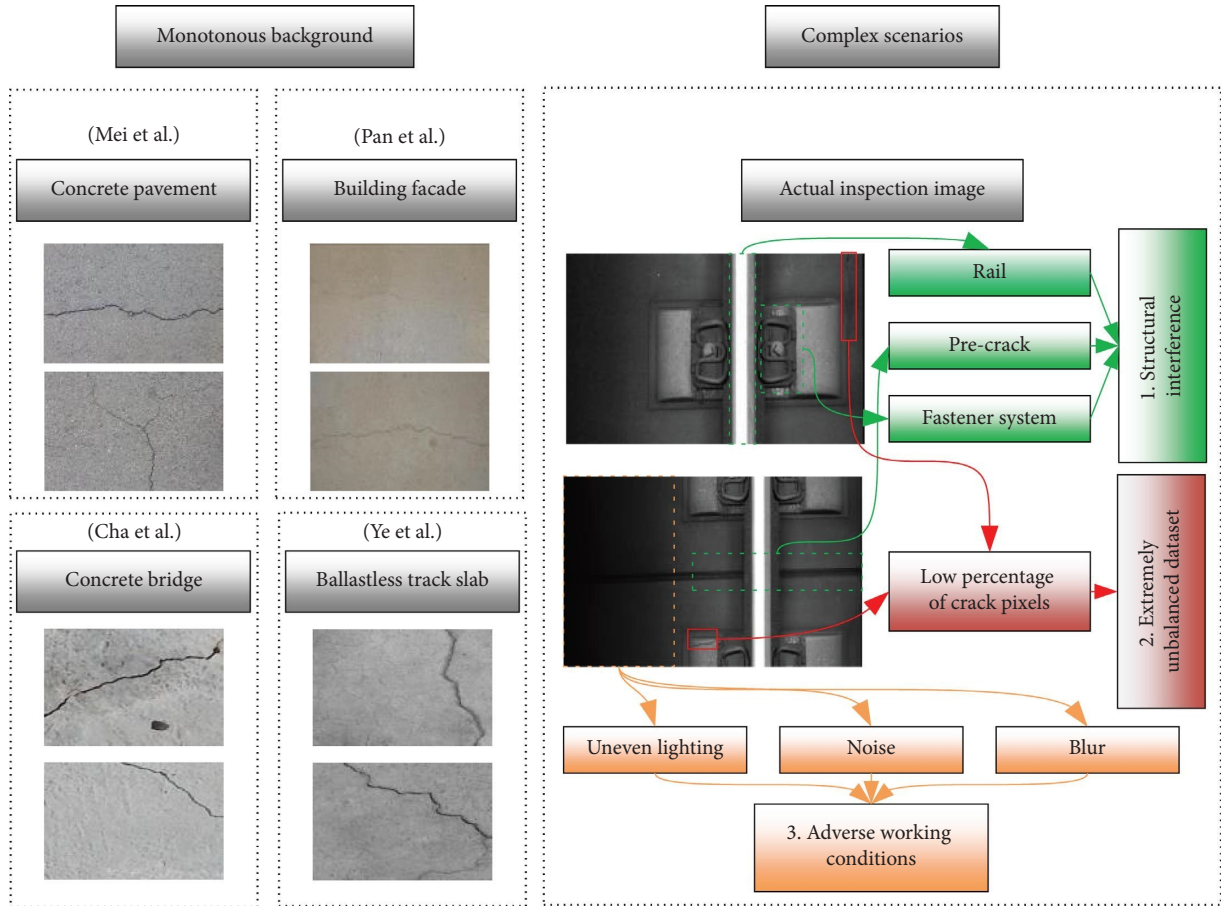


FIGURE 1: Comparison of inspection images of BTS across complex scenarios with generic crack datasets having a monotonous concrete background.

- (3) The inspection for BTS can only be performed at midnight window period, which is not only time-consuming but almost impossible to capture enough valid datasets for training and testing pixel-level deep learning models in the limited time. The preparation of reliable pixel-level ground truth is time-consuming and laborious. Kang et al. [34] noted that it took 30–40 minutes to prepare refined ground truth for each crack image across complex scenarios. Moreover, subjective annotation that relies on the prior knowledge and experience of engineers can lead to low consistency of ground truth.

To address the above challenges, a hybrid algorithm based on synthetic data from digital twin and weakly supervised style transfer is proposed in this paper, which segments cracks accurately and efficiently from inspection images of BTS across complex scenarios. The hybrid algorithm consists of three modules, namely, a synthetic data generation module, a region of concern extraction module, and a weakly supervised crack segmentation module. First, a digital twin model that can simulate real ballast track inspection scenarios is developed, which is used to synthesize crack datasets of BTS across complex scenarios for training deep learning models to

extract the most discriminative features. Then, a region attention-based style transfer algorithm is proposed, which uses a deep object detection network to rapidly screen and locate the region of interest for style transfer, i.e., the crack region, while removing structures and adverse work conditions that interfere in segmentation results. Finally, a cycle-consistent-based generative adversarial network is used for directly transforming weak labels containing crack regions into segmentation results with the same style as the ground truth at the pixel level.

The contributions of the hybrid algorithm proposed in this paper can be summarized as follows:

- (1) The hybrid algorithm proposed in this paper effectively addresses the three challenges posed by inspection images of BTS across complex scenarios. It not only eliminates oversegmentation caused by similar structures such as rails, fastener systems, and precracks thoroughly but also extracts more discriminative features from adverse working conditions such as uneven lighting, noise, and blur for segmenting cracks in a refined manner, compared with existing pixel-level deep learning models.

- (2) An improved style transfer algorithm based on the region attention mechanism is proposed, which focuses the attention of the random, uncontrollable generative adversarial network on small crack regions, overcoming the “all-black image” caused by the low proportion of cracks in inspection images across complex scenarios.
- (3) A synthetic crack dataset that can simulate realistic and complex inspection scenarios of BTS is built for the first time for adequately training deep learning models. The region-based weak supervision is used to replace the pixel-level supervised pattern, which allows pixel-level crack segmentation results to be obtained without using the corresponding ground truth, greatly reducing the time cost of data preparation while overcoming the adverse effects of low consistency of manual annotation.

The remainder of the paper is organized as follows: Section 2 critically reviews the techniques associated with the proposed approach. Section 3 systematically presents the framework and network structure of the proposed hybrid algorithm. Section 4 explains the dataset and evaluation metrics and discusses the experimental results. Section 5 compares and evaluates with existing approaches. Section 6 concludes the paper.

2. Related Works

This section critically reviews the works associated with the proposed hybrid algorithms, including pixel-level deep learning models with customized network structures, physics-based virtual models for generating labeled data, and unsupervised deep learning solutions. The implementation of these related works is dedicated to address three major issues that are faced by existing pixel-level deep learning models when deployed in practice.

2.1. Developing Deep Learning Models with Customized Network Structures. Several researchers have improved the adaptability of deep learning solutions to complex environmental conditions by customizing feature extraction strategies. Zhao et al. [35] proposed a novel crack feature pyramid network (crack-FPN), which exhibited more robust feature extraction capability for crack images affected by lighting conditions and complex backgrounds. Shu et al. [36] integrated the nonforgetting learning method into a 34-layer deep residual network, which avoided feature forgetting in traditional CNNs in processing multitype damage detection for complex structural scenes. Liu et al. [37] proposed a two-stream boundary-aware crack segmentation (BACS) network for high-resolution characterization of cracks against complex backgrounds through the combination of semantic image segmentation and edge detection. Song et al. [38] designed a novel multiscale dilated convolution module that can extract more discriminative crack features under the interference of poor light, noise, and blur. The hierarchical texture-perceiving

generative adversarial network (HTP-GAN) proposed by Gu et al. [39] improved the adaptability to complex environmental conditions by capturing spatially invariant representations of images. The CrackNet-R proposed by Zhang et al. [40] used the mean of a sequence of pixels instead of individual pixel, which can effectively distinguish cracks from noise with higher F-measure. Xiang et al. [41] proposed an improved pixel-level detection model using a combination of channel and spatial attention, which increased the segmentation accuracy by 7% compared with the traditional SOTA model for cracks with complex topological features in low contrast. Zhang et al. [42] designed a ShuttleNet with memory connection to enhance the characterization of asphalt pavement cracks under complex environmental conditions, which obtained 92.54% of F-measure and 86.57% of MIoU.

Although such customized network structures greatly improve the effectiveness and adaptability of pixel-level deep learning models to deployment scenarios, repeated experimentation and tuning based on large amounts of data are indispensable. Whenever the detection scenario changes, the parameters need to be adjusted or even the network structure must be redesigned. Ensemble learning enables combining the scores of multiple deep learning models in a certain way (e.g., fuzzy integrals) to form a final prediction, resulting in better performance than a single model [43]. However, ensemble-based methods still inherit the high specificity of individual models to the detection scenario, which is difficult to obtain satisfying results under the limitations of the image background and training samples. In addition, the low consistency of ground truth based on manual annotation also brings great uncertainty to crack detection by customized models.

2.2. Generating Training Data Based on Physics-Based Virtual Models. Developing physics-based virtual models to generate cracks across complex scenarios is an effective way for driving pixel-level deep learning models to extract more discriminative features when deployed in practice. Hoskere et al. [44] developed parametric finite element models for the infrastructure of multiple sizes and materials, using hot spots in the finite element models to automatically synthesize labeled data for defects such as cracks to adequately train and evaluate deep learning models. The study by Hoskere et al. [45] also pointed out that virtual cracks synthesized based on nonlinear finite element models lead to a 10% improvement in IoU of pixel-level deep learning models compared with using only real data. Pyle et al. [46] used efficient hybrid finite elements (FE) and ray-based simulation to train CNNs for characterizing real cracks in a refined manner. Hakim et al. [47] used 3D finite element model simulation data obtained from a commercial software package to adequately train the neural network for identifying cracks in structures with good adaptability to light. Siu et al. [48] used the game engine to generate sewer pipe damage in virtual environments that can simulate different lighting and camera angles, which allowed an average improvement of 5.8% in AP of faster RCNN compared with using only real data.

However, adequate training on virtual data generated by finite element models does not directly mean good generalization to real data in complex scenarios [49]. Although various finite element models enable accurate simulation of an arbitrary number of various crack types of different material properties with almost no acquisition cost, the difference between virtual and real crack features is also a key factor that results in the reduced robustness of deep learning when deployed in practice. Digital twin models based on physical entities, virtual entities, and the interaction between them enable 3D dynamic perception for infrastructure damage states with a higher degree of simulation [50–52]. The generated data based on digital twin models can not only reproduce the actual damage characteristics realistically but also simulate a variety of structural scenarios compared with finite element models set up in specific boundary conditions. Therefore, a digital twin model integrating real crack features of monotonic scenarios and virtual BIM models of complex scenarios is established in this paper. The generated cracks with real topological features across complex scenarios can minimize the difference between virtual and real data for eliminating the uncertainty of practical deployment.

2.3. Unsupervised Deep Learning Models. Developing unsupervised deep learning-based solutions enables to overcome the low consistency of manual annotation under complex environmental conditions while greatly reducing the cost of data preparation. Convolutional autoencoder (CAE) automatically extracts the compact representations from the input unlabeled images through the encoder module, and only the normal images are reconstructed by the decoder for distinguishing abnormal regions such as cracks [53, 54]. These compact representations are also defined as descriptors by several researchers, and cracks and background are distinguished by comparing the differences in the descriptors learned from each image [55, 56]. Notably, the validity and reliability of these compact descriptors require adequate training by large-scale data and poor for representing fine cracks across complex scenarios.

Generative adversarial networks (GANs) add discriminator structures to judge the effectiveness of image recovery and continuously optimize the parameter settings based on discriminator results, which can extract more compact features to reduce the differences between reconstructed and input images, thus directly transforming cracks into ground truth with similar structural patterns [57–59]. Zhang et al. [60] first employed a cycle-consistent generative adversarial network for unsupervised crack detection, which achieved comparable performance of supervised learning methods without the need of pixel-level ground truth. However, the detection results of such unsupervised GANs are random and uncontrollable, of which they are naturally more concerned with the reconstruction of the whole image. Since the crack region only accounts for a small portion of the whole image, GAN tends to emphasize the reconstruction of background thus outputting an “all-black image.” In addition, backgrounds that are highly similar to cracks such as uneven lighting, random noise, and blur can adversely affect

the distinction between normal and crack regions in image reconstruction, which is a general limitation of unsupervised crack detection.

Therefore, this paper proposes a region attention-based weakly supervised style transfer algorithm, which rapidly captures crack regions in advance using deep object detection networks and uses them as inputs for style transfer. The region-based extraction allows better differentiation between cracks and similar backgrounds and greatly reduces annotation costs compared with segmentation pixel by pixel.

3. Methodology

To address the abovementioned three key issues of existing pixel-level deep learning models in practical deployment under complex environments, this paper proposes a hybrid algorithm of synthetic data by using the digital twin model and weakly supervised style transfer. As shown in Figure 2, the hybrid algorithm consists of three modules: synthetic data generation module, crack region extraction of concern, and weakly supervised crack segmentation. First, the building information model (BIM) for BTS is built, and the lightweight physical engine is used to randomly deploy the collected real BTS cracks of a specific scenario and adverse conditions (uneven lighting, noise, and blur) on the BIM, by which the obtained digital twin model fusing real crack features and virtual BIM is used to synthesize the low-quality rich data required for training deep learning models. Second, a deep object detection network is used to rapidly capture the smallest outer rectangle containing the crack region, which overcomes the “all-black image” while removing the interference of complex backgrounds that are highly similar to cracks. Finally, region-based weak labels are used to train the generative adversarial network for focusing its adversarial loss of attention on the crack regions of interest and directly transferring the crack regions to segmentation results with the same style as the ground truth based on cycle-consistency loss without manual labeling.

3.1. Digital Twin Model-Based Synthetic Data Generation. Digital twin is a technological tool that enables the interaction and integration of the physical and virtual worlds by integrating multiphysical, multiscale, and multidisciplinary attributes, having real-time synchronization, faithful mapping, and high-fidelity characteristics. A digital twin model, defined as a fully parametric, three-dimensional, interactive virtual model built in computer systems, is used to simulate the properties, states, and responses of physical entities in various setting scenarios. In this paper, the digital twin model of BTS is established with the characteristics of real fracture data-driven, virtual-reality interaction and dynamic update. It can reproduce the railway inspection behaviors of real world with high simulation in virtual space to make up for the deficiencies of on-site tests and static BIM. The digital twin model is completely data-driven, which not only realistically reproduces the actual crack features but also simulates any complex inspection scenarios (uneven lighting, background noise, focusing blur, etc.)

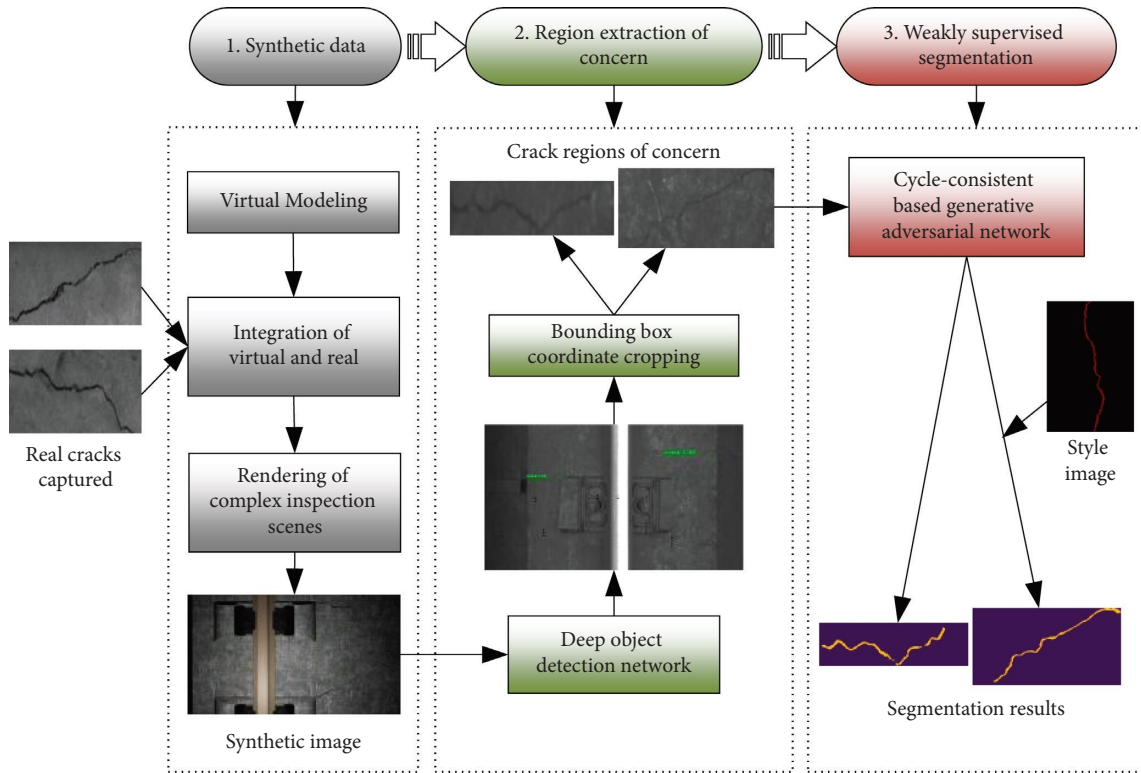


FIGURE 2: Overall framework of the hybrid algorithm.

compared with finite element simulation in a single setup state, thereby providing training data with good generalization for deep learning methods.

Moreover, the digital twin model not only provides the training data required for deep learning methods but also presents an intuitive and visualized dynamic perception platform for inspection results. One-dimensional monitoring data (mechanical response, mileage information, etc.) captured by various static or dynamic sensors can be loaded into the virtual space, and the 2D boundary texture of cracks identified by deep learning methods projected into the digital twin model by UV mapping pixel-by-pixel [61], which are continuously updated with the changes of working conditions and service time, enabling the recording, analyzing, and presenting of the state of health of the BTS throughout its lifecycle. Therefore, the digital twin model established in this paper can dynamically reproduce the service state of BTS in high simulation and continuously update it according to the actual working conditions, compared with the static BIM in the ideal setting state.

As shown in Figure 3, a digital twin model is built to synthesize the training data containing a lightweight BIM, realistic texture and crack features, and virtual inspection scenarios. First, a lightweight BIM for the whole BTS is constructed. Then, the captured realistic crack features and the virtual model are integrated in the physical engine for converting the original BIM into a realistic digital twin model. Finally, the digital twin model is rendered by texturing and lighting for minimizing the differences from the real BTS while obtaining synthetic data that most closely approximates the inspection conditions during the midnight window period.

3.1.1. Digital Twin Model for Portraying BTS Cracking.

This section proposes a lightweight construction solution for railway structure BIM, which can generate 3D BIM on the web side by directly reading the layout rules from 2D CAD drawings using an open element engine, i.e., Three.js, in three steps. Firstly, the center of the inner profile width and centerline of BTS are used as the origin and x -axis, respectively, to establish a plane coordinate system, and the coordinates of the feature points of BTS are determined based on the structural characteristics and interrelationship of each component (rail, fastener system, etc.) of BTS parsed from 2D CAD drawings. Then, the parametric sections of each component are drawn based on the extracted feature points, and the built-in functions are used to perform stretching and sampling for obtaining the parametric components of BTS. Finally, these parametric components are combined in arrays to form the completed BIM of BTS, and step-by-step loading and masking processes are used to achieve a balance between lightweight loading and detailed model presentation. This lightweighting solution not only overcomes the shortcomings of high specificity, low level of automation, and large errors of traditional modeling strategies that rely on manual assembly but also enables to obtain BIM with greater reality and readability.

The constructed virtual BIM is imported into Unreal Engine 5 (UE5) and integrated with captured images of real cracks with monotonous backgrounds for converting the original BIM into a realistic digital twin model. The blueprint editor in UE5 was used to instantiate real crack features of BTS for generating texture subclasses with different levels of detail and graphical parameters. These textures well

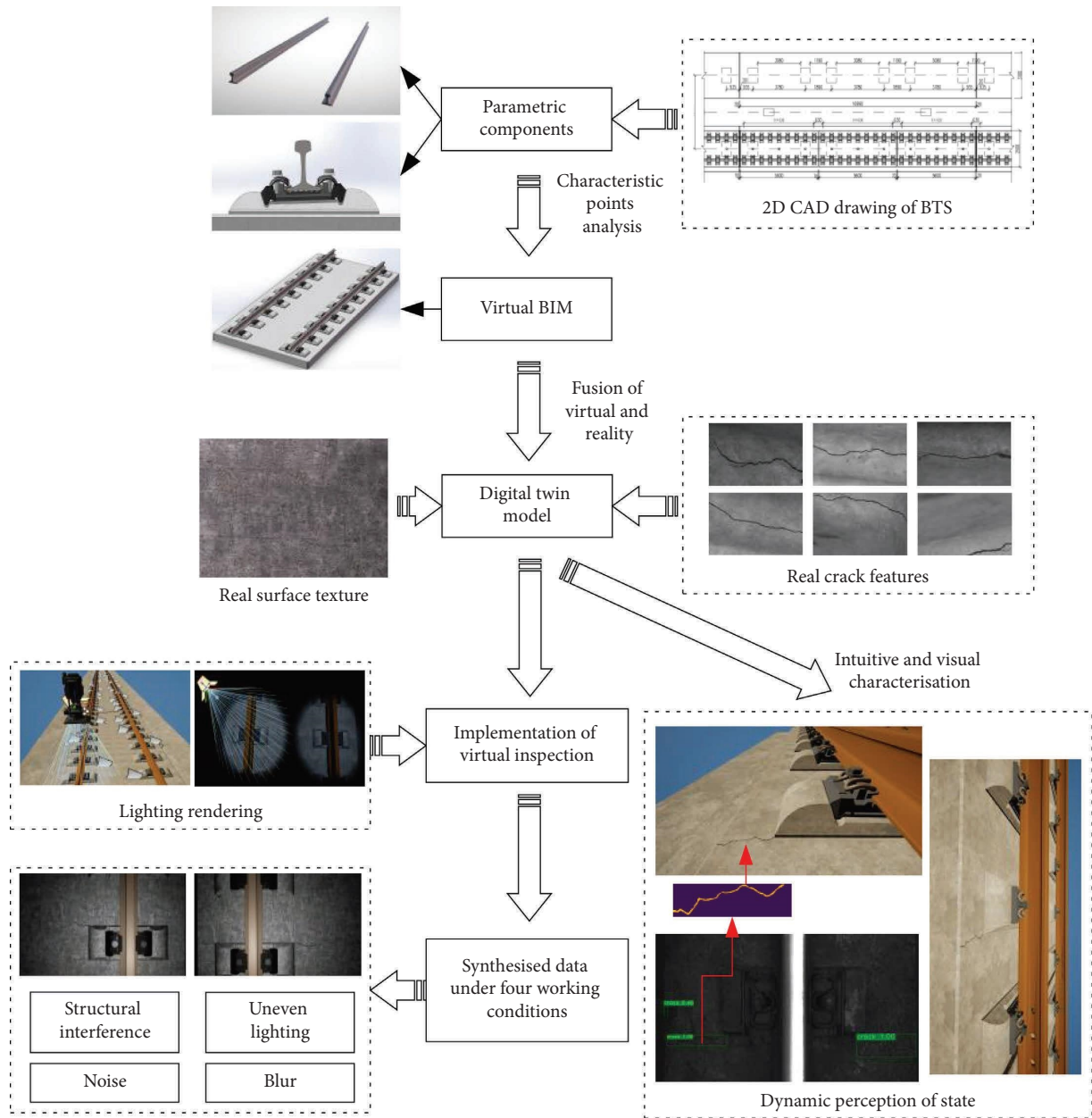


FIGURE 3: Digital twin model-based synthetic dataset preparation.

inherit the geometric topology information of cracks in monotonic scenarios, which are randomly deployed on virtual BIMs while maintaining temporal superresolution for obtaining digital twin models that integrate real crack features and virtual scenarios.

3.1.2. Synthetic Data Simulating Real Inspection Working Conditions. This section performs texture and lighting rendering on the digital twin model of BTS in UE5 for minimizing the difference between the virtual model and the real BTS entity and generating synthetic data that can simulate the real inspection conditions during the midnight window. Firstly, texture maps, bump maps, and reflection maps based on real BTS images are made and attributed to

the digital twin model for simulating the real BTS texture, roughness, and metallic shine while making the lighting interaction on the surface of the digital twin model more realistic. Then, a virtual camera was added in UE5, the line orientation of the virtual camera was calibrated by deploying control points, and lighting assets were added and made to keep pace with the camera movement, which simulates the lighting source installed on the railway inspection vehicle. The range and intensity of the lighting source is adjusted so that the region away from the sides of the rails grows dark for simulating the uneven lighting during the midnight window period. Finally, the salt and pepper noise and Gaussian blur were added to the output of virtual camera for simulating adverse inspection conditions. Salt and pepper noise is implemented by randomly replacing normal pixel points

with black or white noise pixel points in the output synthetic image. A Gaussian function with normal distribution is used to convolve the output image for simulating the blurred image obtained due to camera focus error.

A synthetic crack dataset across complex scenarios is obtained based on the above rendering operations, containing structural interference such as rails, fastener systems, precracks, and adverse inspection conditions with uneven lighting, noise, and blur (Figure 4).

3.2. Crack Region Extraction of Concern. This paper presents a novel region attention-based weakly supervised style transfer scheme, which uses a deep object detection network to precapture the crack regions of interest from the output virtual BTS inspection images. These region-based weak labels are used to focus the attention of random generative adversarial networks on generating the ground truth of the crack pixels, which overcomes the ‘‘all-black image’’ caused by the underrepresentation of crack pixels. A two-stage object detection network, faster R-CNN, is used to extract crack regions of interest, which consists of three components, namely, backbone, region proposal network (RPN), and fast region-based convolutional network (fast R-CNN), as shown in Figure 5. First, multiscale features are extracted from the input virtual inspection images using a feature pyramid network (FPN). Then, these features are used as input to both RPN and fast R-CNN for generating regions of interest. Finally, fast R-CNN is used to classify the region of interest and bounding box regression for distinguishing cracks from the background.

3.2.1. Backbone Network. FPN with a residual network is used as the backbone of faster RCNN, which extracts features with both bottom-level visual information and top-level semantic information from the input image by downsampling, upsampling, and cross-layer fusion. First, a residual network (C2–C5) is used to extract features from input images in a bottom-up manner, and the features obtained from the low level to the high level are used as inputs for successive upsampling. Then, the high-level features are scaled up to the same size as the low-level features by successive upsampling from top to bottom, which are fused with the bottom-up captured features by lateral concatenation for outputting fused feature maps (M2–M5) containing multiscale semantic information. The 1×1 convolutional layer is used to perform a scale-invariant spatial transformation of the features captured by downsampling to accommodate upsampling, and the 3×3 convolutional layer is used to eliminate the feature aliasing due to upsampling for obtaining the feature maps (P2–P5) required by the region proposal network. The general architecture of FPN is shown in Figure 6.

3.2.2. Region Proposal Network. The region proposal network takes the feature maps output by FPN as input and outputs rectangular candidate regions (anchor boxes) of multiple scales and aspect ratios for regions of interest (ROI).

RPN predefines nine benchmark anchor boxes for each sliding window and modifies the benchmark anchor boxes for predicting the proposed region by the four correction parameters obtained from deploying the sliding window on the feature map. As for each input image, the anchor box is marked as a positive sample (crack) if the overlap ratio is greater than 0.7 between the anchor box and the ground truth box and as background if this ratio is less than 0.3.

3.2.3. Fast Region-Based Convolutional Network. Fast R-CNN is used to perform bounding box regression and classification of the proposed regions from RPN. A ROI pooling layer is first used to transform the proposed regions of different shapes and sizes into feature maps of the same size. Then, these feature maps are input to two fully connected layers for bounding box regression and classification, respectively, for predicting the location of the region of interest and determining the class it belongs to. The smooth L1 loss between the prediction box and the labeled box is used as the bounding box regression loss. The crossentropy loss that distinguishes the cracks from the background is used as the classification loss. The loss function of fast R-CNN is shown in the following equation:

$$L = \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \lambda \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}(t_i, t_i^*), \quad (1)$$

where p_i is the probability that the object is included in the prediction box, p_i^* is the labeled box, and t_i and t_i^* are the four parameterized coordinates of the predicted box and the labeled box, respectively.

3.3. Weakly Supervised Crack Segmentation. The proposed region attention-based style transfer only employs weak labels from faster RCNN without manual annotation, which directly converts the crack images into pixel-level segmentation results with the same style as the ground truth of the generic crack forest dataset (CFD). As shown in Figure 7, domain A is derived from the crack region detection results output by faster RCNN, and domain B is derived from the ground truth of CFD. The cycle-consistent-based generative adversarial network consists of two end-to-end GANs that share two generators (G and F) and each takes one discriminator (D_A and D_B). A forward generator G is used to convert A to B , while a reverse generator F is used to convert B to A . The discriminator D_B is used to distinguish the real B from the fake B (\hat{B}) generated by A based on the forward adversarial loss, which encourages the conversion of A to an output indistinguishable from the domain B and vice versa for D_A . The original A is passed through the forward and reverse generators to obtain the reconstructed \hat{A} . The difference between the reconstructed \hat{A} and the original A is defined as a cycle consistency loss for preventing the overfitting of G and F .

3.3.1. Structure of Generators and Discriminators. Generator G is used to convert the input crack image directly into ground truth with the same structural pattern, which consists of

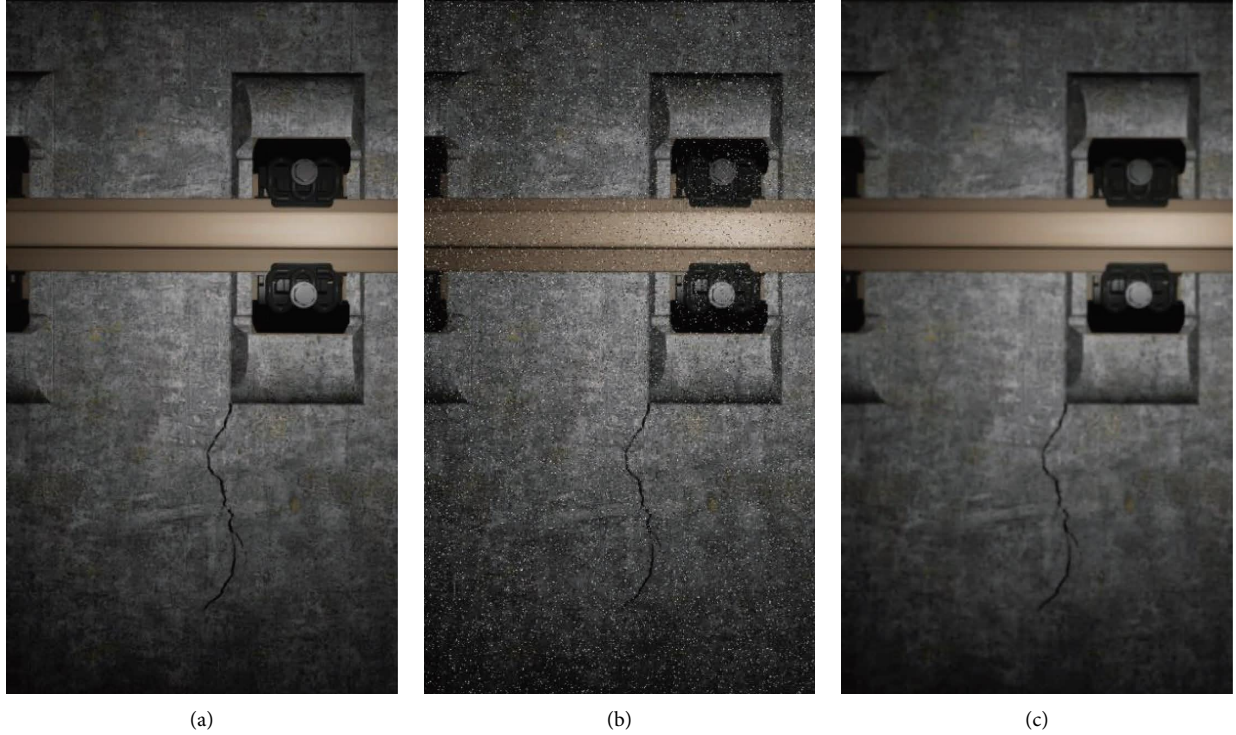


FIGURE 4: Synthetic image with adverse inspection conditions: (a) uneven lighting, (b) noise, and (c) blur.

three modules: encoder, converter, and decoder. There are three convolutional layers used in the encoder to extract features from the input domain A image ($256 \times 256 \times 3$) and compares them into 256 64×64 feature vectors. The converter converts the feature vectors in the domain A to those in the domain B using a 6-layer Resnet module, which enables the style transfer to be performed while preserving the features of the original image. The decoder uses the deconvolution layer to recover the low-level features from the feature vectors output by converter, which generates the fake domain B image, namely, ground truth.

A discriminator consists of five convolutional layers. The four convolutional layers are used to extract features from the input image, which are fed into the convolutional layer

that produces a 1-dimensional output (Decision $[0, 1]$) for determining the classes to which the features belong.

3.3.2. Overall Loss. The goal of the cycle-consistent-based generation adversarial network is to learn two mapping functions (G and F) between domain A (data distribution is $a \sim p_{\text{data}}(a)$) and domain B (data distribution is $b \sim p_{\text{data}}(b)$). The adversarial loss is used to match the data distribution generated by G or F with the real data distribution. The cycle consistency loss is used to prevent G and F from contradicting each other. For the mapping function $G (A \rightarrow B)$ and the discriminator D_B , the adversarial loss is shown in the following equation:

$$L_{\text{GAN}}(G, D_B, A, B) = E_{b \sim p_{\text{data}}(b)}[\log D_B(b)] + E_{a \sim p_{\text{data}}(a)}[\log(1 - D_B(G(a)))]. \quad (2)$$

In addition, the cycle consistency loss is used to ensure that $a \rightarrow G(a) \rightarrow F(G(a)) \approx a$ and $b \rightarrow F(b) \rightarrow G(F(b)) \approx b$, as shown in equation (3). In summary, the total

loss of the cycle-consistent-based generative adversarial network is shown in equation (4).

$$L_{\text{cyc}}(G, F) = E_{a \sim p_{\text{data}}(a)}[\|F(G(a)) - a\|_1] + E_{b \sim p_{\text{data}}(b)}[\|G(F(b)) - b\|_1], \quad (3)$$

$$L(G, F, D_A, D_B) = L_{\text{GAN}}(G, D_B, A, B) + L_{\text{GAN}}(F, D_A, B, A) + \lambda L_{\text{cyc}}(G, F), \quad (4)$$

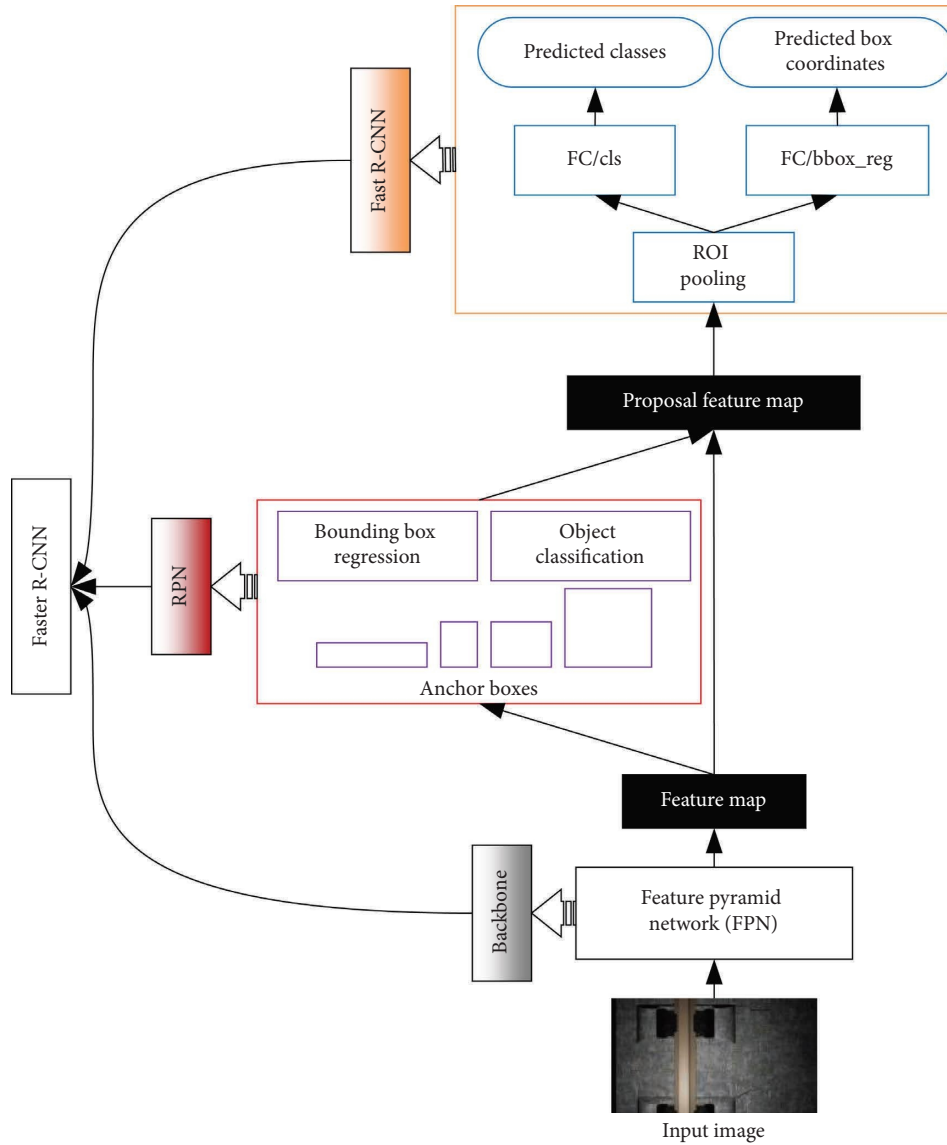


FIGURE 5: Schematic representation of the architecture of faster R-CNN.

where G is to generate the similar $G(a)$ that is indistinguishable from the domain B , while D_B aims to distinguish the generated $G(a)$ from the real domain B . Thus, G and F aim to minimize the objective against the maximization of D_B and D_A , i.e., $\arg \min_{G,F} \max_{D_A,D_B} L(G,F, D_A, D_B)$. λ is used to control the relative importance of these two objectives.

4. Case Study

4.1. Data Preparation. A total of three types of crack datasets with different image backgrounds have been created for verifying and analyzing the performance of the hybrid algorithm: control group (real crack dataset with monotonous background, No. A), training group (synthetic crack dataset across complex scenarios, No. B), and test group (real crack dataset across complex scenarios, No. C).

HD cameras and drones are used to acquire cracks from damaged ballastless track slabs in the high-speed railway laboratory at Central South University (CSU). The real crack images captured by these high-precision devices are characterized by monotonous background, uniform imaging, and high resolution. More than 1000 cropped crack images ($400 \text{ pixels} \times 400 \text{ pixels}$) are selected in this paper for establishing the real crack dataset (control group, No. A) with a monotonous background, of which 200 are used as the testing set and the remaining 800 are used as the training set. In addition, the real topological features contained in these crack images are used to produce texture maps with different detailing properties, which are randomly deployed on the virtual BIM model for generating a digital twin model simulating damage of BTS. Virtual crack images ($1920 \text{ pixels} \times 1080 \text{ pixels}$) across complex backgrounds synthesized from the digital twin model, including structural

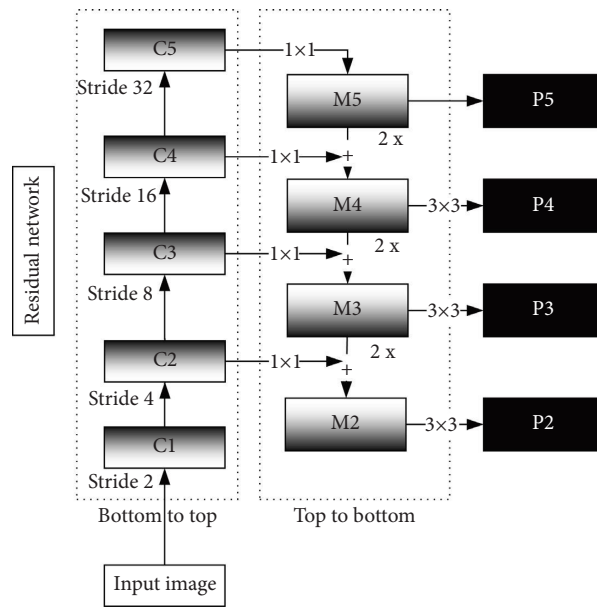


FIGURE 6: Schematic representation of the architecture of FPN.

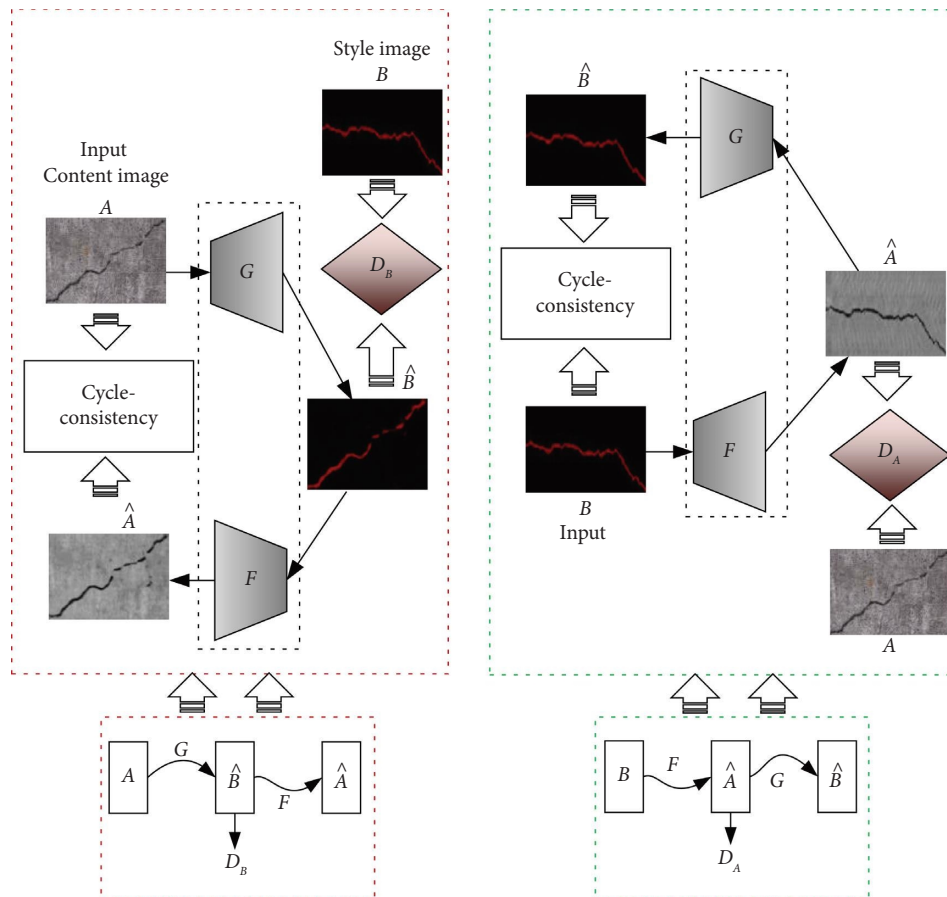


FIGURE 7: Overall structure of the cycle-consistent-based generative adversarial network.

interference, uneven lighting, noise, and blur. These images with complex backgrounds are used to build a synthetic crack dataset (training group, No. B) for training and validating the performance of the hybrid algorithm.

Furthermore, an HD image acquisition system is used to perform on-site experiments for real inspection data, aiming at testing the generalization of the hybrid algorithm to real-world images of cracks. As shown in Figure 8, an electric

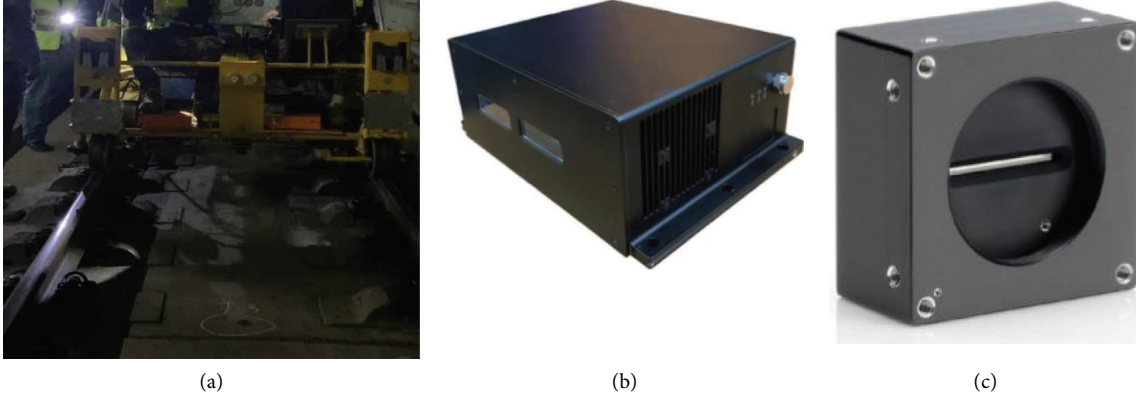


FIGURE 8: Composition of the HD image acquisition system used to perform on-site tests: (a) Electric inspection vehicle. (b) Intelligent inspection host. (c) Image acquisition device.

inspection vehicle, an in-vehicle intelligent inspection host, and an undervehicle image acquisition device are included in the system. The electric inspection vehicle adopts 4-wheel drive design and can travel on the track at a speed of 20 km/h. The intelligent inspection host is used to receive high-definition image data captured from the acquisition end and transfer the results to the control center in real time through the wireless communication module. The image acquisition equipment is 3 sets of 1000 GB Ethernet line array cameras in the bottom part of the vehicle with an infrared strobe laser light source, which can cover the entire BTS surface in a lateral range of 2.4 m.

The output size of the HD image acquisition system is resized to match the virtual data at 1920 pixels \times 1080 pixels, and over 400 crack samples from the real-world BTS are obtained. Among them, 200 images are used as the testing set, and the rest are expanded to 800 images through horizontal flipping and color dithering to be consistent with the scale of the virtual dataset; thus, a real crack dataset across complex scenarios is established (testing group, No. C). Examples and quantities of the three types of crack datasets are shown in Figure 9 and Table 1.

4.2. Performance Evaluation Metrics. The mean of average precision of all classes of detection objects, i.e., MAP is used to evaluate the performance of faster R-CNN. As only one class of detection objects, namely, cracks, is set up in this paper; thus, MAP is AP. The key to MAP calculation is Intersection over Union (IoU), which is defined as the overlap rate between the predicted region and ground truth, and its mathematical expression is shown in the following equation:

$$\text{IoU} = \frac{\text{Predicted region} \cap \text{Ground truth}}{\text{Predicted region} \cup \text{Ground truth}}. \quad (5)$$

The threshold of IoU is generally predefined (set to 0.5 in this paper), and the prediction result is defined as a positive sample when the IoU between the predicted bounding box and the ground truth is greater than this threshold; otherwise, it is a negative sample. In addition, the confidence of the predicted bounding box is also used to distinguish the

true prediction from the false prediction. True positive (TP) is indicated when the IoU of the predicted result is greater than 0.5 and the prediction is true; false positive (FP) is indicated when the IoU of the predicted result is less than 0.5 or the prediction is false; and false negative (FN) is indicated when there is no IoU with ground truth, which indicates that the model cannot detect any object labels from the manual annotation.

Moreover, precision is defined as the ratio of correctly detected objects to the total number of objects detected. Recall is defined as the ratio of correctly detected objects to the total number of real objects. The mathematical expressions for precision and recall are shown in equations (6) and (7). The P-R curve can be plotted by calculating the precision and recall at different confidence thresholds, and the value of MAP is obtained by integrating the P-R curve, which represents the region enclosed by the P-R curve and the coordinate axis.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (6)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (7)$$

The mean of IoU of different classes of detected objects, i.e., MIOU, is used to evaluate the detection results of the cycle-consistent-based generative adversarial network. MIOU is used to measure the overlap rate between the predicted crack pixels belonging to each class and the ground truth, as shown in the following equation:

$$\text{MIOU} = \frac{\text{IoU}}{n}, \quad (8)$$

where n is the class of the detection object.

Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are used to measure the difference between the synthetic images and the real inspection images for assessing the quality of the synthetic crack dataset. PSNR is designed to count the mean square error (MSE) between images, which aims to focus on the differences at the pixel level, as shown in the following equation:

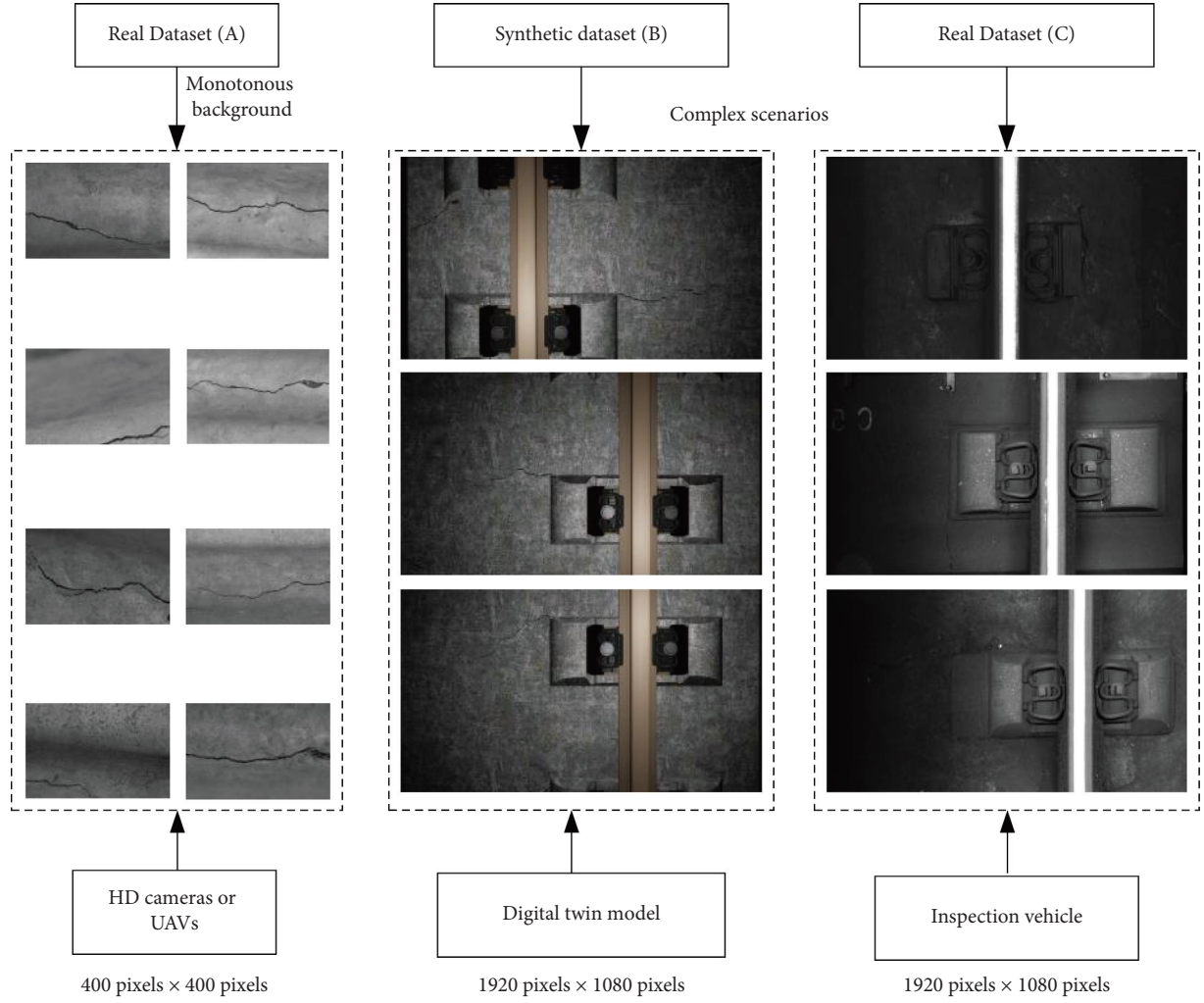


FIGURE 9: Examples of the three types of crack datasets for BTS.

TABLE 1: Composition of the three datasets.

Data type	Real dataset/A	Synthetic dataset/B	Real dataset/C
Characteristics	Monotonous background	Structural interference Uneven lighting Noise Blur	
Image size/pixel	400 × 400	1920 × 1080	1920 × 1080
Training set	800	800	800
Testing set	200	200	200
Total	1000	1000	1000

$$\text{PSNR} = 10 \times \lg \left(\frac{\text{MaxValue}^2}{\text{MSE}} \right), \quad (9)$$

where MSE is the mean square error of the two images. Max Value is the maximum value of the image pixels. The larger the PSNR, the smaller the difference between the synthetic image and the real image, and the better the quality of the synthetic image.

SSIM measures the similarity between two images (x, y) by comparing their lighting $(l(x, y))$, contrast $(c(x, y))$, and structure $(s(x, y))$, as shown in the following equation:

$$\text{SSIM}(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma$$

$$l(x, y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (10)$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3},$$

where μ_x is the mean of x ; μ_y is the mean of y ; σ_x^2 is the variance of x ; σ_y^2 is the variance of y ; σ_{xy} is the covariance of x and y ; and C_1 , C_2 , and C_3 denote three constants to avoid the case where the denominator is zero. The value domain of SSIM is 0 to 1. The larger the SSIM, the higher the structural similarity between the two images.

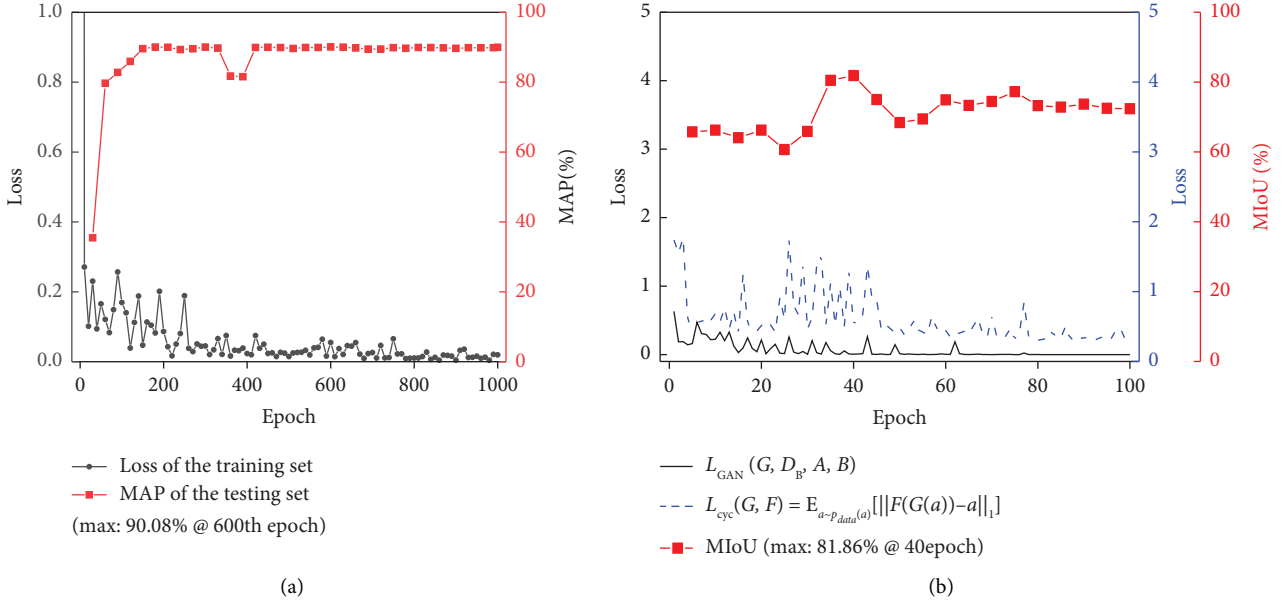


FIGURE 10: Experimental results of the hybrid algorithm: (a) faster R-CNN network; (b) cycle-consistent-based generative adversarial network.

4.3. Experimental Results. The hybrid algorithm for weakly supervised crack segmentation is fully trained and tested on synthetic dataset B using GPU (NVIDIA GeForce RTX 3090) as the computational core and relying on PyTorch 1.2.0, the open-source deep learning framework from Facebook. The optimal hyperparameters of the faster R-CNN network and the cyclic consistent-based generative adversarial network are set as follows: the numbers of epoch are 1000 and 100, respectively; the learning rates are 0.00125 and 0.0002, respectively; both use the adaptive moment estimation algorithm (Adam) to update the weights, and the numbers of images read per update (batch size) are 4.

The training and testing results of the hybrid algorithm on synthetic dataset B using optimal hyperparameters are presented in Figure 10. The total loss of the faster R-CNN network reaches convergence over 1000 epochs of full training, at which time the MAP also reaches stability, indicating that the network has achieved well-fitting state. The best MAP (81.86%) of the faster R-CNN network appears at the 600th epoch, where the output minimum external rectangles containing cracks are used as the input to the cycle-consistent-based generative adversarial network. Both the adversarial loss and the cycle consistency loss of the cycle-consistent-based generative adversarial network reach convergence and stabilization after 100 epochs, indicating that the generator G can excellently convert the output (a) of faster R-CNN network into the segmentation result ($G(a)$) with the same structural pattern as the ground truth (b). The reconstruction results ($F(G(a))$) obtained after two conversions by generators G and F are maximally similar to the output of the faster R-CNN network.

Figure 11 shows the output of each stage of the hybrid algorithm. First, real cracks with the monotone background are randomly deployed in complex inspection scenarios for

synthetic dataset B. Then, the faster R-CNN network is used to capture crack regions accurately and efficiently from inspection images containing complex scenarios. Finally, these weak labels containing cracks are directly converted into segmentation results similar to the ground truth of the CFD dataset. With increasing the number of epochs, the segmentation results show a change from coarse, discontinuous crack features to topologized, continuous crack features (Figure 10), and the optimal crack segmentation results (81.86% MIoU) are reached in the 40th epoch.

5. Discussion

5.1. Experimental Results of Various Detection Algorithms. This section systematically compares the crack segmentation results of the hybrid algorithm, DeepLabv3+ network, and the original cycle-consistent-based generative adversarial network. The MIoU obtained and the time cost required by various algorithms on dataset A (real cracks with the monotonous background) and dataset C (real cracks across complex scenarios) are given in Figure 12. Figure 13 carefully shows the segmentation examples of various algorithms. Although the DeepLabv3+ network achieves the highest MIoU of 85.41% for crack detection in the monotonous background, which drops by over 70% when tested on real crack images across complex scenarios, this indicates that pixel-level deep learning models pretrained using generic datasets designed for general tasks cannot maintain high accuracy on specific tasks across complex scenarios. Structures that are highly similar to cracks, such as pre-cracks, rails, and fastener systems, or even dark image backgrounds, can most adversely affect the generalization and effectiveness of the pretrained model, resulting in catastrophic oversegmentation.

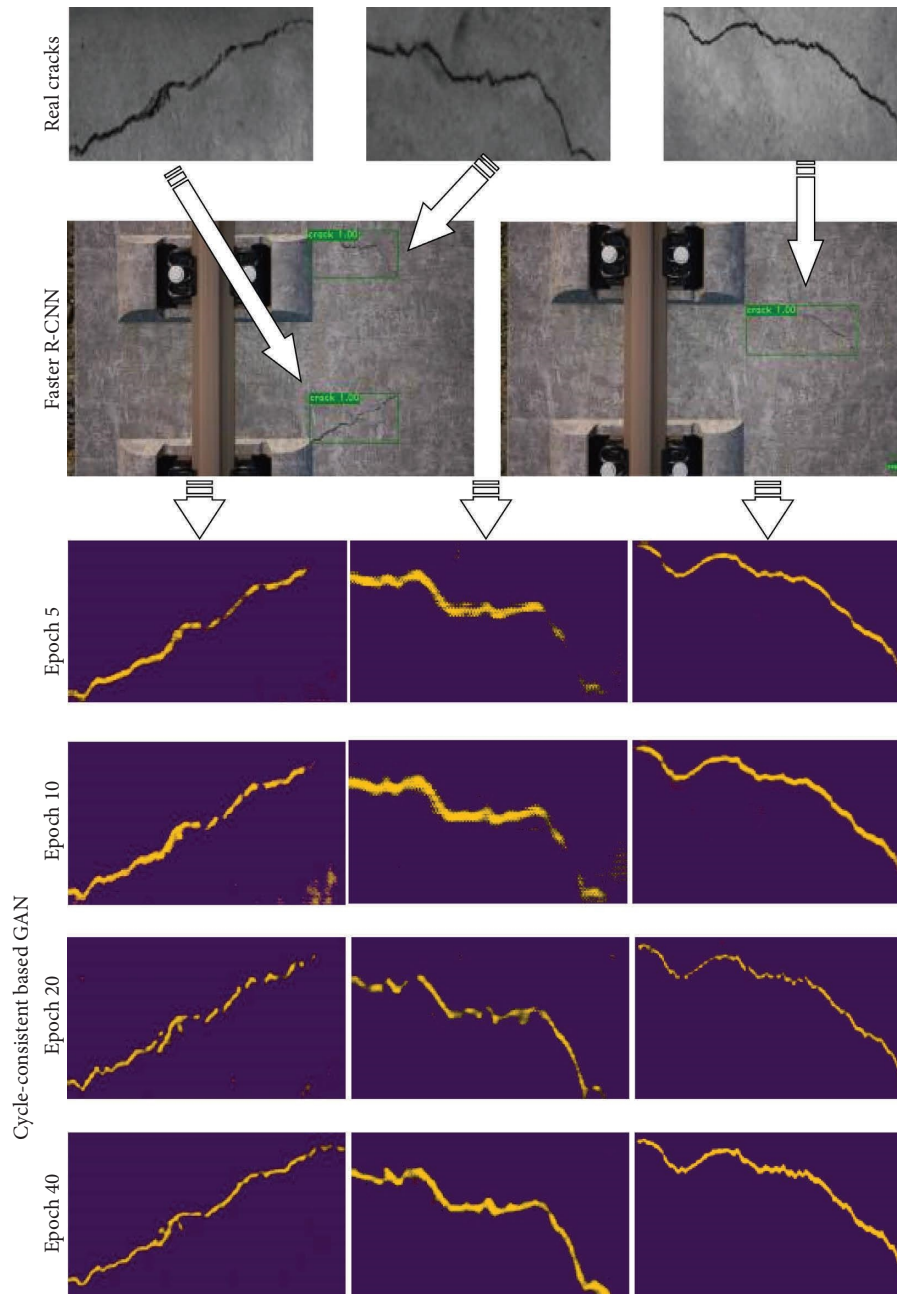


FIGURE 11: Pixel-level crack segmentation results of the hybrid algorithm.

The DeepLabv3+ network, fully trained from synthetic dataset B, achieves an mIoU of over 50% when tested on real crack images across complex scenarios. Also, the over-segmentation caused by irrelevant structures is greatly eliminated and fine-grained crack segmentation results are obtained. This shows that the synthetic dataset created in this paper enables the pixel-level deep learning model to extract more discriminative features than the real crack dataset with the monotonic background, which improves its adaptability to complex scenarios such as structural interference and uneven lighting. Notably, the too low percentage of cracks in

real inspection images still causes the DeepLabv3+ network to generate discontinuous crack boundaries or even “all-black images” leading to missed detection (test results with zero IoU as shown in Figure 13).

A higher mIoU (79.38%) is obtained by the hybrid algorithm on real crack images across complex scenarios, with a nearly 25% improvement compared with the DeepLabv3+ network, producing more fine-grained and continuous crack segmentation results. Particularly, it requires only 0.5% of the annotation time of the DeepLabv3+ network, which overcomes the low consistency of manual annotation pixel

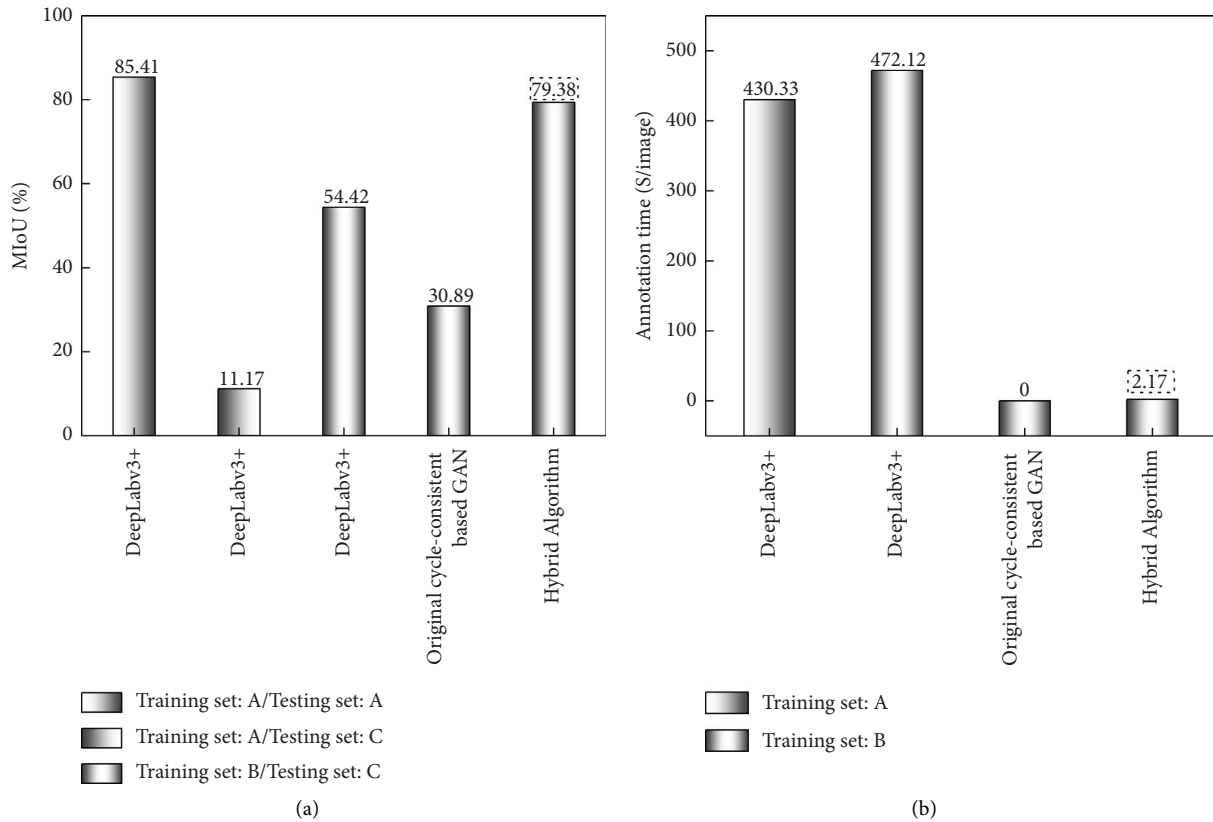


FIGURE 12: MIoU and annotation time of different detection algorithms on real datasets (A and C): (a) MIoU. (b) Annotation time.

by pixel without the need for one-to-one ground truth, demonstrating more accurate and efficient segmentation performance for cracks of BTS.

Furthermore, although the original cycle-consistent-based generative adversarial network can directly convert crack images into segmentation results with the same structural pattern as the ground truth of the CFD dataset without any manual annotation cost, rails, fastener systems, and other ground markings can produce significant oversegmentation, resulting in misdetection. The hybrid algorithm based on region attention enables the original random, uncontrolled generative adversarial network to focus its attention on the crack region of interest, eliminating oversegmentation and obtaining nearly double the MIoU improvement compared with the original GAN.

5.2. Comparison of Training Effectiveness from Synthetic Data and Real Data. The test results on real dataset C of the hybrid algorithm adequately trained based on synthetic dataset B exhibit high MIoU, which is compared with the results of the training relying exclusively on real data in this section, aiming at exploring the alternative and generalization of synthetic data.

Figure 14(a) counts the IoUs on 50 real inspection images (from the testing set in dataset C) of the hybrid algorithm, which are acquired by adequate training from synthetic data

and real data, respectively. More than 70% of all IoUs are obtained by the hybrid algorithm regardless of whether the training set is synthetic or real data. The distribution of the 50 points along the diagnostic line clearly shows that the IoUs obtained from training based on synthetic data are generally close to the training results of real data, with a mean gap of less than 2%. This indicates that the synthetic crack dataset established in this paper enables the performance of the hybrid algorithm to be maximally close to the training effectiveness from the real dataset. In addition, as for the inspection images which are more difficult to identify (IoU is at a lower level, around 70%), the test results of the hybrid algorithm trained based on synthetic data are even better, showing the advantage of the synthetic dataset with rich features over the limited real dataset. Cracks are infrequent, especially since it is difficult to capture enough training samples containing a variety of rich features during the limited midnight window period in a timely manner. Although the size of the training set in real dataset C is expanded by data augmentation (horizon flipping and color dithering) to be consistent with that of synthetic dataset B, the number of effective features that can be used to train deep learning models remains limited. Synthetic data can greatly enlarge the range, type, and number of discriminative crack features compared with data augmentation based on limited real data, thus enhancing the adaptability and robustness of deep learning methods for uncertain inspection data.

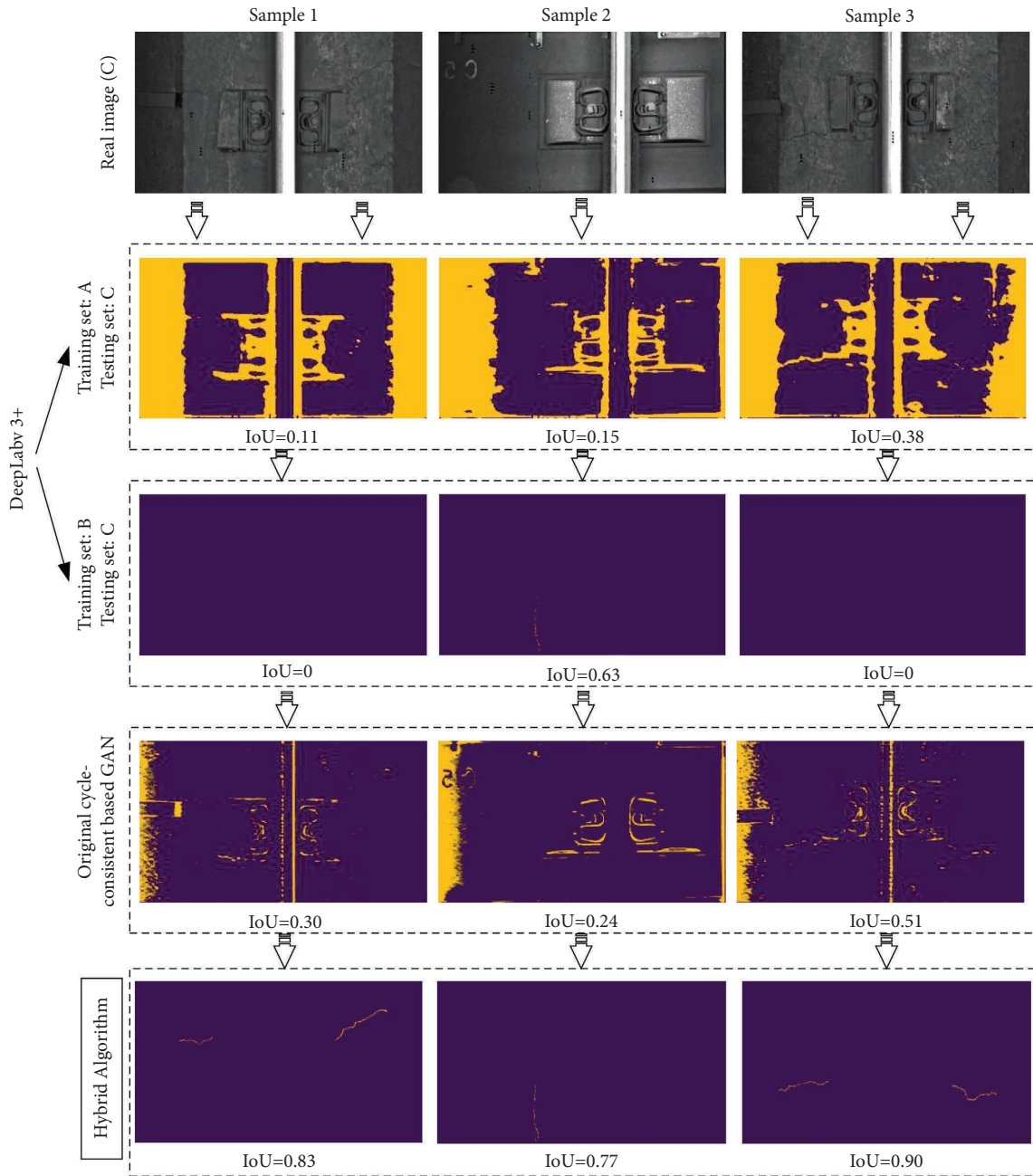


FIGURE 13: Crack segmentation results of different detection algorithms tested on real datasets.

Furthermore, five parallel experiments are implemented to eliminate the uncertainty and randomness of the single training results, and the 50 real inspection images used for testing in each experiment are randomly acquired from the testing set of dataset C. Figure 14(b) counts the relative error rates of the training results based on the synthetic data and those based on the real data. The relative error rates of the training results of the two types of data are below 20% in five parallel experiments, showing a good prospect of synthetic data replacing real data as training samples for deep learning models, enabling a great reduction or even elimination of the labor cost in data acquisition.

Due to differences such as texture between synthetic dataset B and real dataset C, the training results based on synthetic data are slightly lower than those from real data in 80% of cases. Figure 15 quantitatively evaluates the differences between the synthetic and real images at each stage of image generation using two image quality estimators, PSNR and SSIM. The synthetic image rendered with real texture and lighting obtains the highest PSNR and SSIM compared with the virtual image output directly from the BIM, indicating that both have the lowest texture difference at the pixel level in this case. SSIM emphasizes environment perception similar to the human visual system rather than

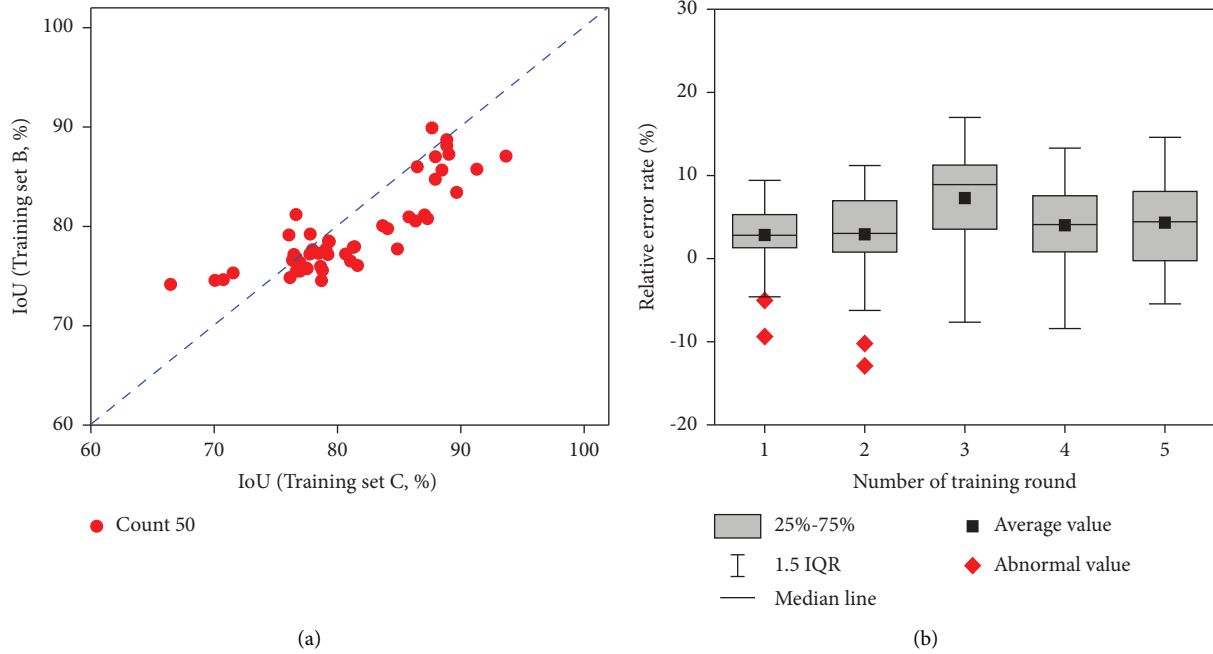


FIGURE 14: Test results over 50 real inspection images by the hybrid algorithm trained from synthetic and real data, respectively: (a) scatterplot of the IoUs; (b) relative error rates of five parallel tests.

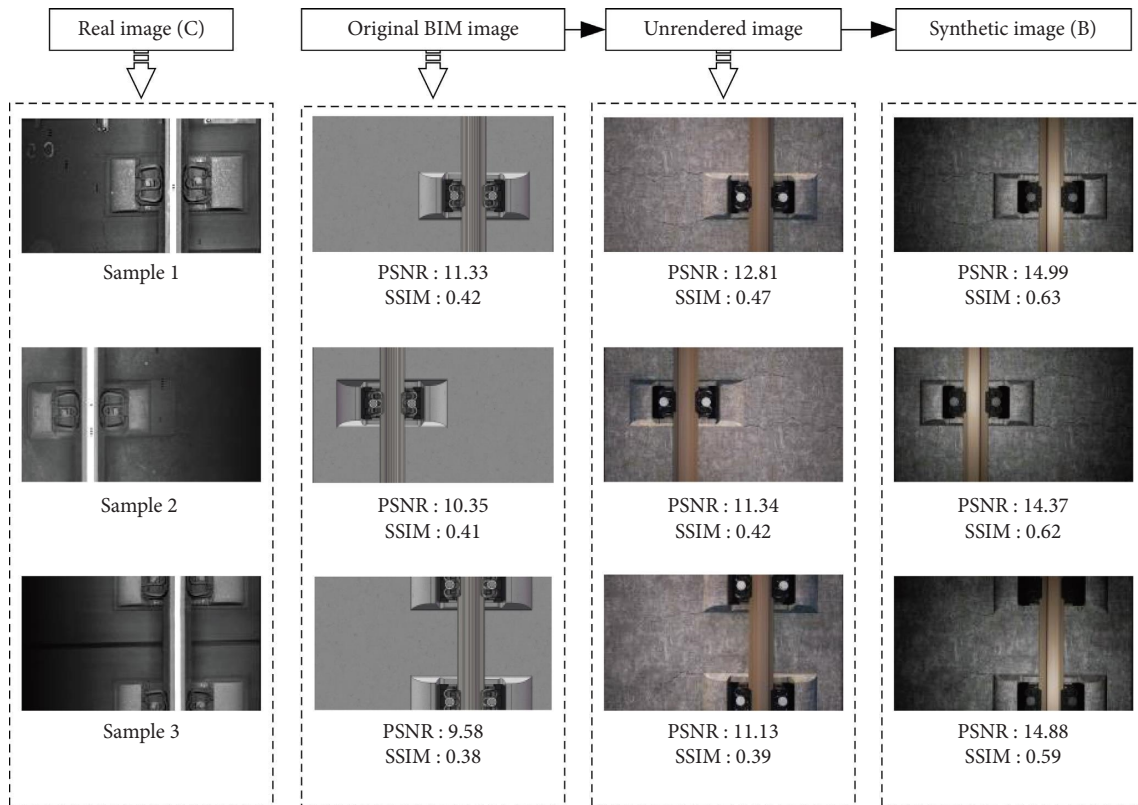


FIGURE 15: Evaluation of differences between synthetic crack images and real inspection images at each stage of image generation.

PSNR, which focuses on differences at the pixel level. The implementation of the virtual inspection with uneven lighting results in an improvement of nearly one-half of the

SSIM between the synthetic image and the real image compared with the original BIM image, which is maximally close to the real inspection working condition during the

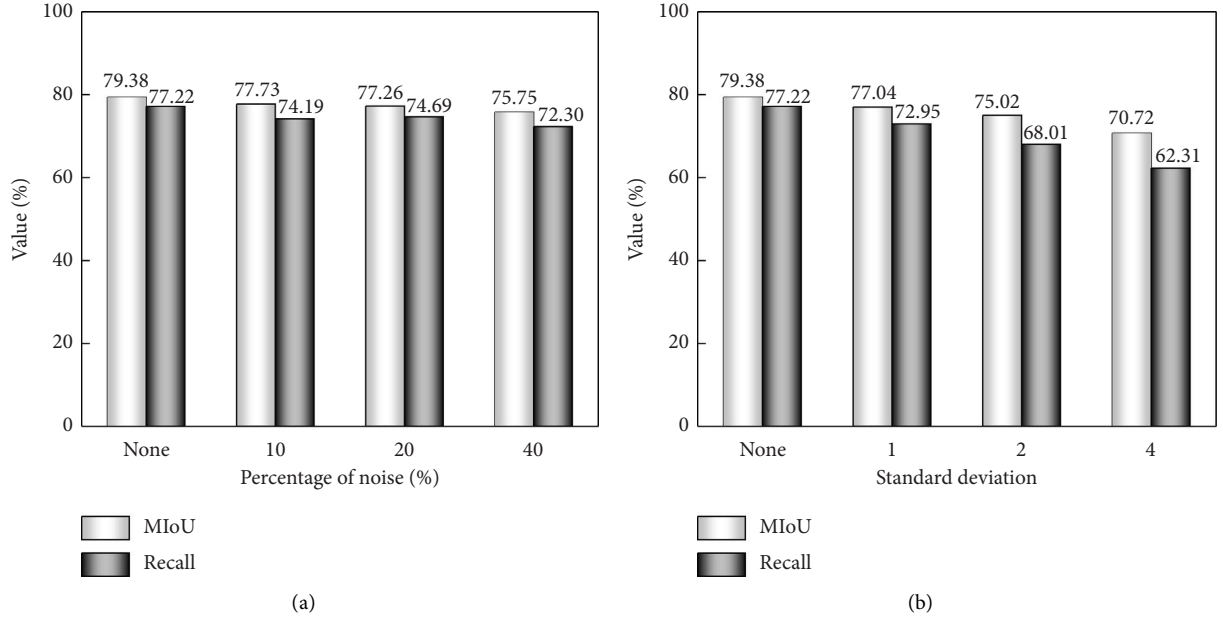


FIGURE 16: Crack detection results of hybrid algorithms under the influence of different proportions of environmental noise and focusing blur. (a) Environmental noise. (b) Focusing blur.

midnight window period. In summary, the continuous reduction of texture and environment differences between the synthetic image and the real inspection image enables the synthetic data to be as close as possible to the real inspection data, which is expected to completely replicate or exceed the training effect of the real data.

5.3. Testing in Adverse Inspection Conditions. The excellent performance of the hybrid algorithm on real inspection images shows strong generalization to structural interference and uneven lighting. Further, the salt and pepper noise and Gaussian blur are added proportionally to 200 test images from the real crack dataset C for analyzing the effect of environmental noise and focusing blur on the performance of the hybrid algorithm. The intensity of environmental noise is simulated by randomly replacing a certain percentage of normal pixel points with white or black noise pixels in inspection images. Gaussian blur uses a Gaussian function with normal distribution to perform a convolution operation on inspection images, which simulates the blurred images acquired due to focusing errors of the image device. The mathematical expression is shown in the following equation:

$$F(r) = \frac{1}{\sqrt{2\pi\sigma^2}N} e^{-r^2/(2\sigma)^2}, \quad (11)$$

where σ is the standard deviation of the normal distribution, the larger the value, the more blurred the image is; r is a Gaussian fuzzy matrix, which is generally taken as $(6\sigma + 1) \times (6\sigma + 1)$ in two-dimensional image space.

As shown in Figure 16, the three proportions of environmental noise have little effect on the crack detection performance of the hybrid algorithm, where MIoU drops within 5% overall. The MIoU of the hybrid algorithm is still above 75% even under the most adverse working conditions, where nearly half of the pixels in the image are converted into noise pixels, showing good adaptability to environmental noise. The drop of MIoU due to focusing blur is twice as large as that of environmental noise, which affects the hybrid algorithm even more adversely. In addition, since omissions are more harmful and of more concern for inspection than errors, the recall of the hybrid algorithm under different working conditions is counted. The drop of recall due to focusing blur is higher at 15% compared with environmental noise, which indicates that blurred images are more likely to cause false-negative errors in the hybrid algorithm, resulting in crack omissions.

The actual segmentation results of the hybrid algorithm under the influence of different proportions of noise and blurring are visualized in Figure 17. Overall, the hybrid algorithm obtains reliable crack segmentation results under various adverse conditions, without “all-black images,” which is far better than the DeepLabv3+ network. With the increasing proportion of noise and blurring, the hybrid algorithm outputs discontinuous crack boundaries and oversegmentation, which is especially serious in blurred images. Therefore, the deployment state of the image devices should be concerned to avoid the adverse effect of blurred images on deep learning methods during the actual inspection of the BTS in the midnight window period.

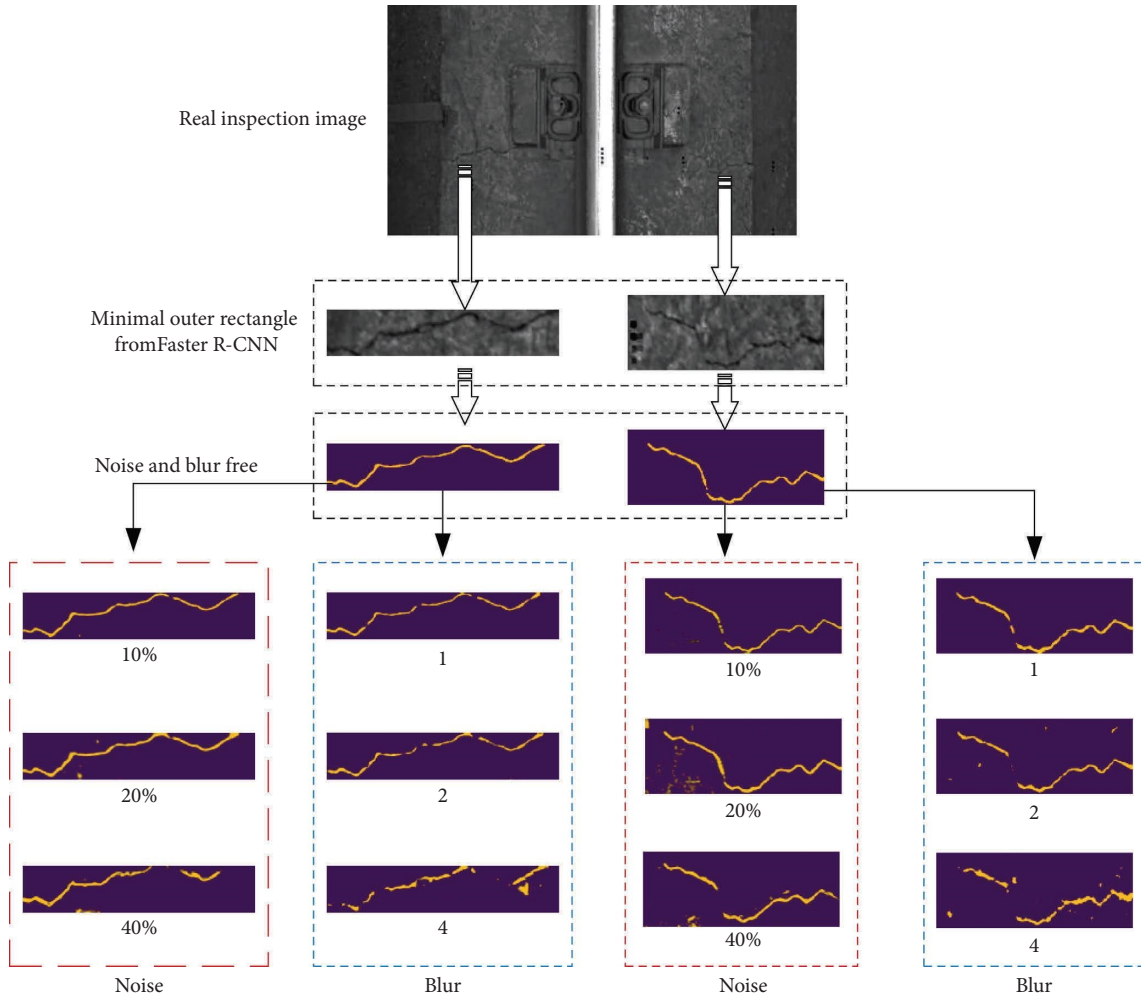


FIGURE 17: Actual segmentation results of the hybrid algorithm under various adverse working conditions.

6. Conclusions

Although the excellent detection performance of existing pixel-level deep learning models for asphalt or concrete pavement cracks with monotonous backgrounds is available, few optimal network structures are available that can segment cracks finely from inspection images of BTS across complex backgrounds. To this end, this paper proposes a hybrid algorithm based on digital twin synthetic data and weakly supervised style transfer, which creatively enables accurate and efficient segmentation of cracks from inspection images containing structural interference such as rails, fastener systems, and precracks, as well as adverse scenarios such as uneven lighting, noise, and blur. The hybrid algorithm achieves the highest MIoU of 79.38% compared with the existing typical pixel-level deep learning model DeepLabv3+ and the original GAN. It not only improves the original coarse oversegmentation or discontinuous few-segmentation into continuous refined segmentation results but also completely overcomes the “all-black image.” In addition, the time cost of the hybrid algorithm is only 0.5% of

that of the DeepLabv3+ network, and pixel-by-pixel crack segmentation is achieved with weak labels at the region level only.

Differences between backgrounds of experimental data used for training and real deployment scenarios may lead to accuracy disasters in deep learning models, which are reviewed for the first time in this paper. The DeepLabv3+ network fully trained with experimental data (similar to the CFD dataset with the monotonic background, uniform lighting, and clear imaging) obtains excellent theoretical accuracy (MIoU of 85.41%), but its performance is only one eighth of the theoretical accuracy when deployed in real inspection scenarios. This indicates that deep learning models fully trained with experimental data (designed for general task requirements) can produce significant slippage in accuracy when deployed directly. Based on this, a synthetic crack dataset of BTS across complex scenarios is created, which aims to simulate the inspection conditions at the midnight window period as realistically as possible and without the cost of acquisition. The IoUs obtained from training based on synthetic data are quite close to the

training results from real data, with an average gap of less than 2%. Synthetic data can greatly enlarge the range, type, and number of discriminative crack features compared with data augmentation based on limited real data, thus enhancing the adaptability and robustness of the hybrid algorithm for uncertain inspection data. Furthermore, the continuous reduction of texture and environment differences between the synthetic image and the real inspection image enables the synthetic data to be as close as possible to the real inspection data, which is expected to completely replicate or exceed the training effect of the real data.

This paper further tests the adaptability and generalization of the hybrid algorithm to various inspection conditions of BTS. The hybrid algorithm obtains reliable crack segmentation results under various adverse conditions, without “all-black images,” which is far better than the DeepLabv3+ network. The MIoU of the hybrid algorithm is still above 75% even under the most adverse working conditions, where nearly half of the pixels in the image are converted into noise pixels. With the increasing proportion of noise and blurring, the hybrid algorithm outputs discontinuous crack boundaries and oversegmentation, which is especially serious in blurred images. Therefore, crack omission from deep learning methods due to blurred images should be avoided when performing the actual inspection of the BTS in the midnight window period.

The hybrid algorithm based on digital twin synthetic data and weakly supervised style transfer proposed in this paper provides an accurate and efficient solution for crack detection in BTS of HSR across complex scenarios. It promotes the advancement from the theoretical experimental level to the practical deployment level of deep learning methods and greatly reduces the cost of data collection and manual annotation. The results greatly improve the reliability and efficiency of maintenance, which is significant to the safe operation of high-speed railway.

Data Availability

Data will be made available on request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. U1734208), the National Natural Science Foundation of China (No. 52178442), and the Hong Kong Polytechnic University’s Postdoc Matching Fund Scheme (No. 1-W29R).

References

- [1] J. Ren, W. Ye, S. Deng, W. Du, and K. Zhang, “Influence of the strain rate on the dynamic damage of cement-asphalt mortar in prefabricated slab tracks,” *Construction and Building Materials*, vol. 299, Article ID 123944, 2021.
- [2] S. Deng, J. Ren, K. Wei, W. Ye, W. Du, and K. Zhang, “Fatigue damage evolution analysis of the CA mortar of ballastless tracks via damage mechanics-finite element full-couple method,” *Construction and Building Materials*, vol. 295, Article ID 123679, 2021.
- [3] J. Ren, S. Deng, K. Wei, H. Li, and J. Wang, “Mechanical property deterioration of the prefabricated concrete slab in mixed passenger and freight railway tracks,” *Construction and Building Materials*, vol. 208, pp. 622–637, 2019.
- [4] J. Wang, X. Z. Liu, and Y. Q. Ni, “A Bayesian probabilistic approach for acoustic emission-based rail condition assessment,” *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 1, pp. 21–34, 2018.
- [5] B. Li, J. Mao, W. Shen, H. Liu, X. Liu, and G. Xu, “Mesoscopic cracking model of cement-based materials subjected to freeze-thaw cycles,” *Construction and Building Materials*, vol. 211, pp. 1050–1064, 2019.
- [6] S. Zhu, M. Wang, W. Zhai et al., “Mechanical property and damage evolution of concrete interface of ballastless track in high-speed railway: experiment and simulation,” *Construction and Building Materials*, vol. 187, pp. 460–473, 2018.
- [7] Z. Zhi-ping, W. Jun-dong, S. Shi-wen, L. Ping, A. A. Shuaibu, and W. Wei-dong, “Experimental study on evolution of mechanical properties of CRTS III ballastless slab track under fatigue load,” *Construction and Building Materials*, vol. 210, pp. 639–649, 2019.
- [8] X. Weng, Y. Huang, and W. Wang, “Segment-based pavement crack quantification,” *Automation in Construction*, vol. 105, Article ID 102819, 2019.
- [9] Y. Zhou, F. Wang, N. Meghanathan, and Y. Huang, “Seed-based approach for automated crack detection from pavement images,” *Transportation Research Record*, vol. 2589, no. 1, pp. 162–171, 2016.
- [10] A. Zhang, Q. Li, K. C. Wang, and S. Qiu, “Matched filtering algorithm for pavement cracking detection,” *Transportation Research Record*, vol. 2367, no. 1, pp. 30–42, 2013.
- [11] I. Abdel-Qader, O. Abudayyeh, and M. E. Kelly, “Analysis of edge-detection techniques for crack identification in bridges,” *Journal of Computing in Civil Engineering*, vol. 17, no. 4, pp. 255–263, 2003.
- [12] H. Zakeri, F. M. Nejad, and A. Fahimifar, “Rahbin: a quadcopter unmanned aerial vehicle based on a systematic image processing approach toward an automated asphalt pavement inspection,” *Automation in Construction*, vol. 72, pp. 211–235, 2016.
- [13] D. Zhang, Q. Li, Y. Chen, M. Cao, L. He, and B. Zhang, “An efficient and reliable coarse-to-fine approach for asphalt pavement crack detection,” *Image and Vision Computing*, vol. 57, pp. 130–146, 2017.
- [14] Y. Tang, X. Zhang, X. Li, and X. Guan, “Application of a new image segmentation method to detection of defects in castings,” *The International Journal of Advanced Manufacturing Technology*, vol. 43, no. 5–6, pp. 431–439, 2009.
- [15] H. Oliveira and P. L. Correia, “Road surface crack detection: improved segmentation with pixel-based refinement,” in *Proceedings of the 2017 25th European Signal Processing Conference (EUSIPCO)*, IEEE, Kos, Greece, August 2017.
- [16] G. Xu, J. Ma, F. Liu, and X. Niu, “Automatic recognition of pavement surface crack based on BP neural network,” in *Proceedings of the 2008 International Conference on Computer and Electrical Engineering*, IEEE, Phuket, Thailand, December 2008.
- [17] H. Oliveira and P. L. Correia, “CrackIT—an image processing toolbox for crack detection and characterization,” in

- Proceedings of the 2014 IEEE international conference on image processing (ICIP)*, IEEE, Paris, France, October 2014.
- [18] Q. Li, Q. Zou, D. Zhang, and Q. Mao, "FoSA: F* seed-growing approach for crack-line detection from pavement images," *Image and Vision Computing*, vol. 29, no. 12, pp. 861–872, 2011.
 - [19] Y. Hou, Q. Li, C. Zhang et al., "The state-of-the-art review on applications of intrusive sensing, image processing techniques, and machine learning methods in pavement monitoring and analysis," *Engineering*, vol. 7, no. 6, pp. 845–856, 2021.
 - [20] H. Maeda, Y. Sekimoto, T. Seto, T. Kashiyama, and H. Omata, "Road damage detection and classification using deep neural networks with smartphone images," *Computer-Aided Civil and Infrastructure Engineering*, vol. 33, no. 12, pp. 1127–1141, 2018.
 - [21] P. Pan, Y. Xu, C. Xing, and Y. Chen, "Crack detection for nuclear containments based on multi-feature fused semantic segmentation," *Construction and Building Materials*, vol. 329, Article ID 127137, 2022.
 - [22] Y. Ren, J. Huang, Z. Hong et al., "Image-based concrete crack detection in tunnels using deep fully convolutional networks," *Construction and Building Materials*, vol. 234, Article ID 117367, 2020.
 - [23] W. Wang, W. Hu, W. Wang et al., "Automated crack severity level detection and classification for ballastless track slab using deep convolutional neural network," *Automation in Construction*, vol. 124, Article ID 103484, 2021.
 - [24] J. Wang, X. He, S. Faming, G. Lu, H. Cong, and Q. Jiang, "A real-time bridge crack detection method based on an improved inception-resnet-v2 structure," *IEEE Access*, vol. 9, pp. 93209–93223, 2021.
 - [25] S. Dorafshan, R. J. Thomas, and M. Maguire, "Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete," *Construction and Building Materials*, vol. 186, pp. 1031–1045, 2018.
 - [26] Z. Liu, Y. Cao, Y. Wang, and W. Wang, "Computer vision-based concrete crack detection using U-net fully convolutional networks," *Automation in Construction*, vol. 104, pp. 129–139, 2019.
 - [27] A. Zhang, K. C. Wang, B. Li et al., "Automated pixel-level pavement crack detection on 3D asphalt surfaces using a deep-learning network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 10, pp. 805–819, 2017.
 - [28] W. Ding, H. Yang, K. Yu, and J. Shu, "Crack detection and quantification for concrete structures using UAV and transformer," *Automation in Construction*, vol. 152, Article ID 104929, 2023.
 - [29] Q. Mei and M. Gül, "A cost effective solution for pavement crack inspection using cameras and deep neural networks," *Construction and Building Materials*, vol. 256, Article ID 119397, 2020.
 - [30] Y. Pan, G. Zhang, and L. Zhang, "A spatial-channel hierarchical deep learning network for pixel-level automated crack detection," *Automation in Construction*, vol. 119, Article ID 103357, 2020.
 - [31] Y. J. Cha, W. Choi, and O. Büyüköztürk, "Deep learning-based crack damage detection using convolutional neural networks," *Computer-Aided Civil and Infrastructure Engineering*, vol. 32, no. 5, pp. 361–378, 2017.
 - [32] W. Ye, S. Deng, J. Ren, X. Xu, K. Zhang, and W. Du, "Deep learning-based fast detection of apparent concrete crack in slab tracks with dilated convolution," *Construction and Building Materials*, vol. 329, Article ID 127157, 2022.
 - [33] J. Deng, A. Singh, Y. Zhou, Y. Lu, and V. C. S. Lee, "Review on computer vision-based crack detection and quantification methodologies for civil structures," *Construction and Building Materials*, vol. 356, Article ID 129238, 2022.
 - [34] D. Kang, S. S. Benipal, D. L. Gopal, and Y. J. Cha, "Hybrid pixel-level concrete crack segmentation and quantification across complex backgrounds using deep learning," *Automation in Construction*, vol. 118, Article ID 103291, 2020.
 - [35] W. Zhao, Y. Liu, J. Zhang, Y. Shao, and J. Shu, "Automatic pixel-level crack detection and evaluation of concrete structures using deep learning," *Structural Control and Health Monitoring*, vol. 29, no. 8, 2022.
 - [36] J. Shu, W. Ding, J. Zhang, F. Lin, and Y. Duan, "Continual-learning-based framework for structural damage recognition," *Structural Control and Health Monitoring*, vol. 29, no. 11, 2022.
 - [37] G. Liu, W. Ding, J. Shu, A. Strauss, and Y. Duan, "Two-stream boundary-aware neural network for concrete crack segmentation and quantification," *Structural Control and Health Monitoring*, vol. 2023, Article ID 3301106, 17 pages, 2023.
 - [38] W. Song, G. Jia, H. Zhu, D. Jia, and L. Gao, "Automated pavement crack damage detection using deep multiscale convolutional features," *Journal of Advanced Transportation*, vol. 2020, Article ID 6412562, 11 pages, 2020.
 - [39] L. Gu, L. Zhang, and Z. Wang, "A one-shot texture-perceiving generative adversarial network for unsupervised surface inspection," in *Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, Anchorage, AK, USA, September 2021.
 - [40] A. Zhang, K. C. Wang, Y. Fei et al., "Automated pixel-level pavement crack detection on 3D asphalt surfaces with a recurrent neural network," *Computer-Aided Civil and Infrastructure Engineering*, vol. 34, no. 3, pp. 213–229, 2019.
 - [41] X. Xiang, Y. Zhang, and A. El Saddik, "Pavement crack detection network based on pyramid structure and attention mechanism," *IET Image Processing*, vol. 14, no. 8, pp. 1580–1586, 2020.
 - [42] A. A. Zhang, K. C. Wang, Y. Liu et al., "Intelligent pixel-level detection of multiple distresses and surface design features on asphalt pavements," *Computer-Aided Civil and Infrastructure Engineering*, vol. 37, no. 13, pp. 1654–1673, 2022.
 - [43] G. Cyganov, A. Rychenkov, A. Sinitca, and D. Kaplun, "Using the fuzzy integrals for the ensemble-based segmentation of asphalt cracks," *Industrial Artificial Intelligence*, vol. 1, no. 1, p. 5, 2023.
 - [44] V. Hoskere, Y. Narazaki, B. F. Spencer, and M. D. Smith, "Deep learning-based damage detection of miter gates using synthetic imagery from computer graphics," in *Proceedings of the 12th International Workshop on Structural Health Monitoring: Enabling Intelligent Life-Cycle Health Management for Industry Internet of Things (IIOT)*, IWSHM 2019, DEStech Publications Inc, Stanford, CA, USA, September 2019.
 - [45] V. Hoskere, Y. Narazaki, and B. F. Spencer, "Learning to detect important visual changes for structural inspections using physicsbased graphics models," in *Proceedings of the 9th International Conference on Structural Health Monitoring of Intelligent Infrastructure: Transferring Research into Practice, SHMII 2019*, International Society for Structural Health Monitoring of Intelligent, St. Louis, MO, USA, August 2019.
 - [46] R. J. Pyle, R. L. Bevan, R. R. Hughes, R. K. Rachev, A. A. S. Ali, and P. D. Wilcox, "Deep learning for ultrasonic crack characterization in NDE," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 68, no. 5, pp. 1854–1865, 2021.

- [47] S. J. S. Hakim, H. Abdul Razak, and S. A. Ravanfar, "Fault diagnosis on beam-like structures from modal parameters using artificial neural networks," *Measurement*, vol. 76, pp. 45–61, 2015.
- [48] C. Siu, M. Wang, and J. C. Cheng, "A framework for synthetic image generation and augmentation for improving automatic sewer pipe defect detection," *Automation in Construction*, vol. 137, Article ID 104213, 2022.
- [49] W. Hu, W. Wang, C. Ai et al., "Machine vision-based surface crack analysis for transportation infrastructure," *Automation in Construction*, vol. 132, Article ID 103973, 2021.
- [50] F. Jiang, L. Ma, T. Broyd, and K. Chen, "Digital twin and its implementations in the civil engineering sector," *Automation in Construction*, vol. 130, Article ID 103838, 2021.
- [51] T. G. Ritto and F. A. Rochinha, "Digital twin, physics-based model, and machine learning applied to damage detection in structures," *Mechanical Systems and Signal Processing*, vol. 155, Article ID 107614, 2021.
- [52] W. Wang, X. Xu, J. Peng, W. Hu, and D. Wu, "Fine-grained detection of pavement distress based on integrated data using digital twin," *Applied Sciences*, vol. 13, no. 7, p. 4549, 2023.
- [53] J. K. Chow, Z. Su, J. Wu, P. S. Tan, X. Mao, and Y. H. Wang, "Anomaly detection of defects on concrete structures with the convolutional autoencoder," *Advanced Engineering Informatics*, vol. 45, Article ID 101105, 2020.
- [54] K. Lee, S. Jeong, S. H. Sim, and D. H. Shin, "A novelty detection approach for tendons of prestressed concrete bridges based on a convolutional autoencoder and acceleration data," *Sensors*, vol. 19, no. 7, p. 1633, 2019.
- [55] Z. Rastin, G. Ghodrati Amiri, and E. Darvishan, "Unsupervised structural damage detection technique based on a deep convolutional autoencoder," *Shock and Vibration*, vol. 2021, Article ID 6658575, 11 pages, 2021.
- [56] A. Mujeeb, W. Dai, M. Erdt, and A. Sourin, "Unsupervised surface defect detection using deep autoencoders and data augmentation," in *Proceedings of the 2018 International Conference on Cyberworlds (CW)*, IEEE, Singapore, October 2018.
- [57] Z. Gao, B. Peng, T. Li, and C. Gou, "Generative adversarial networks for road crack image segmentation," in *Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN)*, IEEE, Budapest, Hungary, July 2019.
- [58] K. Zhang, Y. Zhang, and H. D. Cheng, "CrackGAN: pavement crack detection using partially accurate ground truths based on generative adversarial learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 2, pp. 1306–1319, 2021.
- [59] Y. Yan, S. Zhu, S. Ma, Y. Guo, and Z. Yu, "CycleADC-Net: a crack segmentation method based on multi-scale feature fusion," *Measurement*, vol. 204, Article ID 112107, 2022.
- [60] K. Zhang, Y. Zhang, and H. D. Cheng, "Self-supervised structure learning for crack detection based on cycle-consistent generative adversarial networks," *Journal of Computing in Civil Engineering*, vol. 34, no. 3, Article ID 04020004, 2020.
- [61] W. Yang, Z. Chen, C. Chen, G. Chen, and K. Y. K. Wong, "Deep face video inpainting via UV mapping," *IEEE Transactions on Image Processing*, vol. 32, pp. 1145–1157, 2023.