An Automated Inspection Method for the Steel Box Girder Bottom of Long-Span Bridges Based on Deep Learning

DALEI WANG, YIQUAN ZHANG, YUE PAN, BO PENG, HAORAN LIU, AND RUJIN MA

1College of Civil Engineering, Tongji University, Shanghai 200092, China
2College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
3Division of Engineering, New York University Abu Dhabi, Abu Dhabi, UAE

Corresponding author: Yue Pan (1410001@tongji.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 51878493.

ABSTRACT Among the existing methods for the maintenance and monitoring of bridges, human eye evaluation, which is inevitably subjective and time-consuming, is still the most widely applied. In this paper, a new automatic inspection method for the deterioration of the bottom of a steel box girder based on computer vision is proposed. First, a computer vision system installed on a bridge inspection vehicle is used to capture photos of the bottom of the steel box girder, which are synthesized into panoramas by image stitching technology. Then, a U-net based semantic segmentation network is used to identify the diseases in the panoramas. Finally, the statistics of the disease are executed to evaluate the health condition of the box girder bottom. Comparisons between various sets of deep neural network models are also carried out. Our experimental results show that this method is effective and feasible as a replacement for manual inspection and can achieve a more standardized and accurate evaluation. This method has great engineering potential value in the progress of intelligent structural health management, and should be extended to solve other similar problems.

INDEX TERMS Bridge engineering, deep learning, image stitching, disease detection.

I. INTRODUCTION

As key components of road systems, bridges play an extremely important role in civil engineering. The structural performance of a bridge will inevitably degrade with increases in working time, and may cause serious traffic accidents if degradation is not detected in a timely manner. Therefore, regular detection plays a significant role in bridge management and maintenance. Traditionally, deterioration of a bridge is detected by manual measurements made with portable instruments, in order to obtain the key information of the structure for subsequent analysis and diagnosis. However, this method is unreliable and inefficient, as it heavily depends on human experience. In recent years, with the rapid development of computer technology, intelligent detection has been proposed as a more effective method. Many scholars have made great contributions to the hardware and algorithms towards making it become the mainstream approach in the future.

The factors that cause structural diseases are complex. At present, there are no widely accepted quantitative evaluation criteria, which leads to the issue of subjectivity in detection. Studies have shown that detection methods based on computer vision can reduce or even avoid such subjective errors. According to the technical methods used, the detection of bridge structural diseases based on computer vision can be divided into image feature recognition methods and deep learning detection methods.

In early stages, image features were relatively easier to extract by applying image processing methods. Tsao and Chen first combined an expert system and image features to achieve the classification and detection of apparent concrete multi-class diseases [1]. Based on fuzzy set theory, the image was filtered and pavement cracks were extracted by binarization and clustering. With the development of computer hardware and the support vector machine (SVM) method, Jahanshahi et al. used an artificial neural network (ANN)
and SVM to classify and extract fractures [2]. Prasanna et al. noticed the drawbacks of traditional methods, such as the random forest (RF) algorithm and SVM [3]. Therefore RF, SVM, and AdaBoost classifiers were trained, evaluated, and compared by extracting multi-scale features to achieve concrete crack recognition, and structural apparent panoramas were synthesized and disease density maps were constructed by image stitching. The results showed that the AdaBoost classifier performed the best, achieving an accuracy of more than 90%. In order to improve the performance of SVM, Li et al. utilized a novel feature selection approach based on the linear SVM with a greedy search strategy, detected the fracture area of the contour image extracted by an iterated Canny operator, and then eliminated non-fracture noise [4].

In recent years, with the rapid development of computer technology, deep learning has been gradually integrated into disease detection, which has drawn the interest of many scholars. At present, detection methods based on deep learning can be divided into three categories: networks based on image classification, object detection, and semantic segmentation.

Image classification is an effective method for disease detection. Cha et al. applied convolutional neural networks (CNNs) to disease detection and classification [5]. Their test results showed that this method not only has good adaptability to adverse factors such as uneven light and shadow, but also performs well in distinguishing the differences between shadows and cracks. In addition, Xu et al. proposed a fatigue crack identification method for steel box girders based on a classification network and a sliding window [6]. Their method is made up of three steps: First, the sliding window method is used to cut the original image. Second, a deep neural network is constructed to achieve the classification of three categories of steel box girder cracks, handwriting traces, and background. Finally, the extraction of crack contours is achieved by the trained neural network. The proposed network had the ability to classify cracks and handwriting traces, but had poor accuracy on high-resolution images. Comparing the characteristics of traditional image processing methods and deep convolutional neural network methods, Dorafshan et al. proposed a concrete crack detection method combining the AlexNet classification network and log edge detection [7]. In the same period, Atha et al. also compared the recognition ability of different sizes of sliding windows and different classification networks for detection of apparent corrosion in steel structures [8].

Object detection is different from image classification, due to the fact that it can directly output the specific location of the identified target in a panoramic image, rather than just distinguishing the categories of the image. Cha et al. accomplished the detection and recognition of structural bolt corrosion, concrete cracks, steel member corrosion, and steel structure apparent deterioration diseases, based on a fast R-CNN network [9]. The experimental results showed that the average accuracy of multi-category detection reached 87.8%, and only took 0.03 seconds on a 500 px × 375 px image under the support of a high-performance GPU, thus nearly achieving real-time detection. Combining migration learning and CNNs, Dung et al. achieved the detection of fatigue crack areas in a steel structure [10]. The detection accuracy reached 98% in their own test set, but the method cannot achieve real-time detection.

Compared with classification and detection networks, segmentation networks can directly obtain the contour and specific location of a target object. Zhang et al. proposed an apparent crack segmentation network, named CrackNet, and proved its accuracy and effectiveness through experiments [11]. Considering that CrackNet has no pooling layer (to downsize the outputs of previous layers), Dung et al. proposed a crack extraction method based on a fully convolutional network (FCN), which could achieve more than 90% detection accuracy on linear cracks. For circular concrete peeling, the proposed network also showed good performance [12]. Based on creating a new layer of image feature fusion on FCN, Liu et al. proposed a concrete crack recognition technology based on the U-Net model [13]. Their results showed that the detection effect of the U-Net network was better than that of a target detection network.

In this paper, an automated inspection method for the bottoms of steel box girders in long-span bridges, based on deep learning technology, is proposed. The main contributions of this paper are as follows:

1) A novel automated inspection method for the bottom of steel box girders in long-span bridges is proposed, including image acquisition, image stitching, and a semantic segmentation network for disease recognition.

2) By comparing the performances of different famous neural network architectures, the U-Net with a VGG-16 backbone is found to be the best architecture for steel box girder bottom coating deterioration identification.

3) Based on the vision data, a standardized structural health evaluation method is proposed, which is able to illustrate the health condition distribution of box girder bottom over the whole bridge, and is more effective and accurate for the further qualitative and quantitative analysis.

The remainder of this paper is organized as follows. The image acquisition system will be introduced in Section 2. Section 3 will introduce the image stitching technology. The basic details of the proposed neural network—the semantic segmentation and recognition network for steel structure apparent diseases—will be introduced in Section 4. Then, experiments carried out on the Jiangyin Bridge, Jiangsu Province are illustrated in detail in Section 5. Finally, in Chapter 6, our conclusions, a summary of our limitations, and expectations for future work are discussed.
II. IMAGE ACQUISITION SYSTEM

A. IMAGING SYSTEM

The imaging system we used was composed of an industrial camera and lenses, and a slide rail. The industrial camera and lenses were used to take images of the beam bottom, while the slide rail was applied to fix the camera and reduce the impact of vibration of the steel box girder on imaging. The whole hardware system was arranged on an inspection vehicle at the bottom of steel box girder, as shown in Figure 1. The industrial camera used was a Basler acA2040-120uc camera with 20–60 mm lenses. Some of its core parameters are shown in Table 1. The slide rail was composed of two guide rails, a platform, supports, and other components; wherein the guide rails were directly fixed to the inspection car and were stabilized through the supports. The camera platform was placed on the guide rail to fix the industrial camera, and the platform could slide on the guide rails to adjust the shooting position of the camera.

B. DISTORTION CORRECTING

Lens distortion is the general term for the inherent perspective distortion of an optical lens, which means the distortion caused by the optical structure and imaging characteristics of the lens. Lens distortion is difficult to eliminate by physical means. Image distortion is caused by the inconsistency of local magnification in the visual field, which affects the image quality of taken photos. Common distortion are as follows. Barrel distortion and pillow distortion are caused by the symmetrical radial distortion induced by the characteristics of lens shape, while centrifugal distortion is caused by asymmetric radial distortion and tangential distortion due to the lens not being completely parallel to the image plane. Affine distortion is caused by inconsistency of the length and width of photographic elements. Generally, a distortion correction algorithm is applied to solve for the distortion model parameters and achieve correction. In close-range imaging, the most commonly used distortion model is the Brown–Conrady model [14], which is defined as follows:

For symmetrical radial distortion:

$$\begin{align*}
\delta r &= k_1 r^3 + k_2 r^5 + k_3 r^7 + \ldots \\
\delta x_{\text{rad}} &= \frac{x - x_0}{r} \delta r \\
\delta y_{\text{rad}} &= \frac{y - y_0}{r} \delta r.
\end{align*}$$

(1)

For centrifugal distortion:

$$\begin{align*}
\delta x_{\text{tan}} &= (1 - \frac{f}{U}) [p_1 r^2 + 2(x - x_0)] + 2p_2(x - x_0)(y - y_0) \\
\delta y_{\text{tan}} &= (1 - \frac{f}{U}) [p_2 r^2 + 2(y - y_0)] + 2p_1(x - x_0)(y - y_0).
\end{align*}$$

(2)

For affine distortion:

$$\begin{align*}
\delta x_{\text{aff}} &= b_1 x + b_2 y \\
\delta y_{\text{aff}} &= 0.
\end{align*}$$

(3)

As a result, the gross distortion of a lens can be described as:

$$\begin{align*}
\Delta x &= (x - x_0)(k_1 r^2 + k_2 r^4 + k_3 r^6 + \ldots) + p_1 r^2 + 2(x - x_0) + 2p_2(x - x_0)(y - y_0) + b_1(x - x_0) + b_2(y - y_0) \\
\Delta y &= (y - y_0)(k_1 r^2 + k_2 r^4 + k_3 r^6 + \ldots) + p_2 r^2 + 2(y - y_0) + 2p_1(x - x_0)(y - y_0) + r = \sqrt{(x - x_0)^2 + (y - y_0)^2},
\end{align*}$$

(4)

where $x_0, y_0, f$ are the X and Y co-ordinates of the image’s main point and the camera focal length, respectively; $U$ is the object distance; $k_i$ are the radial distortion parameters of different orders; $p_i$ are centrifugal distortion parameters; and $b_i$ are the affine distortion parameters.

C. IMAGE QUALITY CONTROLS

In order to reduce perspective distortion and ensure a controllable imaging range and accuracy, the shooting angle in Figure 2 should be as small as possible. If the imaging co-ordinates of two points $A$ and $B$ on the scanned surface are $I_A$ and $I_B$, respectively, then according to the trigonometric geometric relationship, the actual co-ordinates $X_A$ and $X_B$ satisfy the following relationship:

$$\begin{align*}
X_A &= \frac{U I_A \mu}{f \cos^2 \theta - I_A \mu \cos \theta \sin \theta}, \\
X_B &= \frac{U I_B \mu}{f \cos^2 \theta - I_B \mu \cos \theta \sin \theta},
\end{align*}$$

(5)

where $\mu$ is the pixel size.
According to equation 5, considering the actual size of a single pixel when $\theta$ is small, the actual size $S_P$ can be simplified as:

$$S_P = \frac{X_A - X_B}{I_A - I_B} \approx \frac{U}{f \cos^2 \theta \mu}.$$  \hfill (6)

In practical applications, in order to reduce the cost of hardware, lenses with different focal lengths are often changed to meet task requirements under different conditions. The focal length $f$ needs to satisfy:

$$f \geq \frac{L\mu}{S_P + \mu},$$  \hfill (7)

where $L$ is the observation distance, and $S_P$ is the actual pixel size (i.e., the observation accuracy).

Comparing equations 6 and 7:

$$f \geq \frac{L\mu}{S_P \cos^2 \theta + \mu}.$$  \hfill (8)

Equation 8 can effectively help to select the appropriate lens focal length and camera specifications, according to the shooting distance and shooting angle, in order to meet the requirements of structural apparent diseases scanning.

In order to reconstruct a panoramic image of the scanned surface of the structure, the collected images need to reach a certain overlap rate to meet the requirements of image stitching. The purpose of setting up the scanning motion control standard was to select reasonable parameters, such as scanning movement rate, sampling frequency, exposure time, and image overlap rate, in the process of image acquisition. The schematic diagram of image stitching under motion control is shown in Figure 3. In order to ensure that there was an overlapping area with an overlapping rate of $\phi$ between successive images, when the driving platform moves at a speed $v$ parallel to the scanned surface, a reasonable sampling frequency can also be selected according to formula 9.

$$fps \geq \frac{v}{nS_P(1-\phi)} \approx \frac{f \cos^2 \theta}{nU \mu(1-\phi)}k,$$  \hfill (9)

where $fps$ is the sampling frequency of the camera, $n$ is the number of pixels in each image, and $k$ is an calculation coefficient with a value of 1.2.

It should be noted that the sampling frequency of the camera has no relationship with the shutter speed. In the process of motion, if the exposure time is too long, the image will be dragged. Therefore, it is necessary to select a shutter speed with sufficient exposure and no drag, according to the lighting conditions and aperture size. The estimation formula is as follows:

$$t \leq \frac{S_P}{v}.$$  \hfill (10)

According to equations 9 and 10, it can be inferred that if $0.1$ mm is used as the observation precision to collect $1920 \times 1280$ px images, with an overlap rate of $50\%$, when the moving rate of the driving platform is $0.1$m/s along the long edge of the image, the sampling frequency of the machine must be at least $1.25$ frames/s, which requires taking an image with a length of $1920$ pixels every $0.8$ seconds. At the same time, if the pixel size is $3 \mu m$, in order to ensure that the image does not suffer from dragging, the exposure time should be $30$ ms at most.

III. IMAGE STITCHING

Image stitching originates from the field of photogrammetry. In early stages, fine homologous image points between ground control points and points on images were established manually to connect the geometric transformation relationship among multiple aerial photos to complete image registration and, finally, to synthesize large-scale stitched images. The steps in image stitching include sequence image acquisition, image pre-processing, image registration, image transformation, and image fusion. Among these steps, image registration is the key to the success of image stitching.

A. IMAGE REGISTRATION

With the emergence and development of digital images, in order to achieve automatic stitching, many registration algorithms based on calculating image information have been proposed to replace manual annotation. In this research, feature-based techniques are mostly applied.

The idea of feature-based image stitching is to extract feature information from the image, then to recognize corresponding feature areas by matching features between two images, followed by rectifying and merging the images to complete stitching. The key step in this type of method is the extraction of image feature information. Commonly used feature information include contour features, gradient features, point features, and image moments. At present, feature point matching-based methods are mainly used. Classic point feature methods include Harris et al. [15], KLT [16], Shi [17], SUSAN [18], FAST [19], ORB [20], SIFT [21], and SURF [22]. Among them, the SIFT point feature proposed by David G. Lowe is the most robust and widely used, due to its invariance to rotation, scaling, brightness changes, and so on.
B. BUNDLE ADJUSTMENT

For long-sequence image stitching, due to parallax, pixel error, shooting angle error, and other reasons, the conversion relationship obtained by registration between the same points of different images only achieves optimal local matching, but not globally. Therefore, when an image sequence is too long, errors will accumulate and drift will occur. In order to reduce such error, it is necessary to optimize the registration error through the bundle method after the registration of feature points; namely:

\[ e_{\text{min}} = \min_{i} \sum_{j \in I(i)} \sum_{k \in P(i,j)} d(T(P_{j,i}^{k}), T(P_{j,i}^k, t_{k}))^2, \quad (11) \]

where \( i \) is the image sequence; \( I(i) \) is the image set matching with image \( i \); \( T(P_{j,i}^k) \) is the point that \( P_j \) is transformed into after transformation \( t_{k} \); and \( d \) is the Euclidean distance.

When the internal parameters \((x_0, y_0, f)\) of a long sequence of pictures taken by the same camera remain unchanged, it is only necessary to determine the external parameters \((R_{3 \times 3}, T)\) of the camera for each shot, in order to establish the homography relationship between the associated images and obtain the homography matrix for global error minimization.

C. IMAGE FUSION

After image stitching, there are obvious stitching gaps (or ghosts) in the overlapped area of the image, due to imaging differences or the existence of moving objects. Therefore, the overlapped parts of the images need to be fused. Image fusion technology determines the final image synthesis quality. In this study, a multi-stage fusion method is used.

The basic idea of a multi-stage fusion method is to decompose the image into images of different frequencies for superposition. For different frequencies, the weighted method is used for fusion: For the low-frequency part, a weighted signal with a wide wavelength (for example, \( \sigma \) in the Gaussian kernel function is relatively large) is used; and a narrower signal is used for the higher frequency part.

The specific calculation method is as follows: first, construct the corresponding Laplacian pyramids \( L_t \) and \( L_{t+1} \) for the overlapping area of the two images—generally using the method similar to SIFT scale space construction—to build the DoG image pyramid \( L(x, y) \). Then, the Laplacian pyramid at the same level is weighted and fused to obtain the fused Laplacian pyramid. Finally, the pyramid is extended from the top and summed with the image of the next layer. Expand in turn, until the bottom of the pyramid is restored.

At present, the multi-stage fusion method is still a good option. The stitched image is clear, smooth, and seamless, which can avoid the problems of gaps and overlaps. Therefore, if enough computing power is available, it is better to use this method to obtain high-quality fused images.

IV. SEMANTIC SEGMENTATION NETWORK FOR RECOGNITION OF STEEL STRUCTURE DISEASES

In recent years, with the rapid development of the Internet economy, big data has become easier to collect and the ability for humans to obtain and process information has been greatly improved. In particular, as high-performance calculating chips technologies have been developed and promoted, it has become possible to train and use deep neural networks with more layers and parameters, which has been accomplished by a new generation of artificial intelligence technologies, represented by so-called “deep learning”. At present, various technologies based on deep learning have made remarkable breakthroughs in image comprehension, speech recognition, machine translation, auto-driving, and other fields.

Constructing convolutional neural networks is the key for deep learning technology to accomplish computer vision tasks. From the proposal of the first convolutional network, for the recognition of handwritten digits at the end of 20th century, a large variety of convolutional neural networks of different structures have been developed. Especially after 2012, many scholars have experimentally proven their network structures to be effective. Among these, the more famous ones include AlexNet [23], ZFNet [24], GoogleNet (InceptionV1 [25], V2 [26], V3 [27], V4 [28]), VGG [29], ResNet [30], DenseNet [31], Xception [32], ResNeXt [33], PolyNet [34], DPN [35], SENet [36], and NASNet [37] (with the ability to automatically design its own network structure of itself); as well as lightweight networks, including SqueezeNet [38], ShuffleNet [39], and MobileNets V1 [40], V2 [41], and V3 [42].

The recognition of steel structure apparent deterioration by computer vision methods can be regarded as a semantic segmentation task [43], [44]. Its basic steps are to construct a fully convolutional network, down-sample the image by part of the network and acquire a low-resolution feature map, up-sample the feature map using another part of the network, and finally output a partition template with every pixel annotated by class. At present, the networks that are commonly used for such purposes include SegNet [45], U-Net [46], LinkNet [47], DeepLab [48], [49], RefineNet [50], Enet [51], and so on.

A. CONVOLUTIONAL NEURAL NETWORK

Based on the mechanism of biological vision, the idea of a convolutional neural network (CNN) was first proposed and applied to image recognition in the 1980s. To simulate the structure of the visual cortex in the human brain, a convolutional neural network can be constructed by stacking a series of specific network layers. The basic network layers includes convolutional, pooling, activation, batch standardization, and deconvolutional layers, among others.

1) CONVOLUTIONAL LAYER

A Convolutional layer is the core component of an image segmentation network. The name “convolutional neural
network” refers to adding a convolution operation into a traditional neural network to form a “convolutional layer”, in order to realize certain functions by the neural network. The intuition of convolution is that, for a given input image and convolution template, let the convolution template slide over the input image, extract the pixel values in the central area of the convolution template and multiply them with the template every time, and then sum all the values as the value of the output image at this location. The essence of convolution is the linear superposition of each point \((x, y)\) and its surrounding points in the input image. The superposition weight depends on the convolution template and, so, convolution can also be understood as a kind of local statistical information of the image. The calculation process in a convolutional neural network acts to extract the local statistical information of the input image and to integrate the information into more abstract “features”. The standard expression of multi-channel convolution is:

\[
I'(x, y, k) = \sum_{i=-p}^{p} \sum_{j=-q}^{q} \sum_{c=1}^{C} I(x+i, y+j, c)K_c(i, j, k),
\]

where \(I(x, y, c)\) is the value of the original image at the position of \((x, y)\) in channel \(k\); \(I'(i, j, k)\) is the value of the output image at \((x, y)\) in channel \(k\); and \(K_c(i, j, k)\) is the value of the convolution template mapping to channel \(k\) (also called the “kernel”) at \((x, y)\) in channel \(c\).

2) DECONVOLUTIONAL LAYER
For an image segmentation network, it is generally necessary to up-sample the feature map after down-sampling to obtain a segmentation template which has the same resolution as the input image. Deconvolution is similar to convolution, using a template to slide on the input image to obtain the corresponding output value. The difference is that deconvolution needs to up-sample the input image once before the actual operation, double the resolution of the input image, and then carry out the ordinary convolution operation. The calculation method is shown in Figure 4.

B. SEMANTIC SEGMENTATION FULLY CONVOLUTIONAL NETWORK ARCHITECTURE
Considering that the surface features of a steel structure are not complex and that network efficiency is required, due to the huge amount of structural scanning data, this paper proposes an effective fully convolutional semantic segmentation network as the segmentation network for apparent deterioration area detection of steel structure, in order to achieve rapid disease identification.

The segmentation network is divided into two parts: an encoder (down-sampling) and a decoder (up-sampling). In the process of sampling on the network, the feature map of the up-sampling and the corresponding size of the down-sampling feature map are integrated to improve the network’s ability to express features. We trained multiple networks in an attempt to achieve the best performance. Among these, the best architecture is shown in Figure 5. The input layer size of the architecture is \(128 \times 128\) 3-channel images and the output is a \(128 \times 128\) 1-channel template. The value of each template pixel represents the confidence of the category. In the process of down-sampling, the image resolution is reduced to \(4 \times 4\) by stacking five consecutive convolutional modules, as well as pooling layers. After that, using one deconvolutional layer, a feature image fusion layer, and one or two convolutional layers as the basic units, the image is up-sampled and the image resolution is restored to \(128 \times 128\).

C. LOSS FUNCTION
In order to learn the parameters in each network layer, it is necessary to design a reasonable loss function to calculate the difference between the current network inference results and the actual values when training the network. Then, the optimal parameters can be obtained from the whole network in an iterative process using an optimization algorithm. Generally, the smaller the loss function, the better the fitting degree of the model. According to different task objectives, loss functions are often different. However, in practical applications, the generalization ability of a model must be considered to prevent over-fitting; thus, it is not always true that smaller loss means better results. In order to minimize the empirical
and structural risks, the actual objective function is:

$$\tilde{J}(\theta) = \arg\min \frac{1}{N} \sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(\theta),$$

(13)

where $y_i$ is the true value; $f(x_i)$ is the prediction of input $x_i$ given by the model; $\theta$ is the weight vector of the model; $J(\theta)$ is the regularization term; and $\lambda$ is regularization coefficient.

In practice, the cross-entropy loss function is often used:

$$\tilde{J}(\theta) = J(\theta) + \lambda \Omega(\theta)$$

$$= \sum_{i=1}^{m} (-y^{(i)} \ln p - (1 - y^{(i)}) \ln(1-p)) + \frac{\lambda}{2} \theta^T \theta,$$

(14)

where $\lambda$ is set to 0.01.

V. EXPERIMENT
A. GENERAL INFORMATION

In this section, an experiment on the JiangYin Bridge, Jiangsu Province, China is introduced. The Jiangyin Yangtze River Highway Bridge, referred to as “Jiangyin Bridge”, is a river passage connecting Taizhou City and Wuxi City in Jiangsu Province, China, which is located on the Yangtze River waterway and was opened to traffic on September 28, 1999. The total length of the Jiangyin Bridge is 3071 meters, and the main bridge is 1385 meters. The bridge deck is a two-way six-lane expressway, with a design speed of 100 km/h. It is a suspension bridge with double towers and double cable steel box girders. The main span is a steel box girder with air nozzles, and the two sides are pre-stressed concrete continuous box girders with the same height as the main span. The north bank approach bridge is composed of pre-stressed concrete simple supported girders, while the south is composed of right mountain road and pre-stressed concrete viaducts. The main beam is a flat streamline steel box girder. The steel box girder panel is an orthotropic steel box girder, the sling anchor box is set outside the air nozzle, and the sling tension is transmitted to the web and diaphragm through the hanging plate and three stiffening force transfer plates.

Our image acquisition system was installed on the bridge inspection vehicle at the bottom of the steel box girder. It scanned the bottom of the main girder of the Jiangyin bridge, from the south end to the north end, on two different strips, as shown in Figure 6.
Table 2. Distortion coefficients.

| Distortion coefficients | Value     |
|-------------------------|-----------|
| k1                      | -1.095    |
| k2                      | 10.325    |
| k3                      | -200.578  |
| k4                      | 2.979 \times 10^{-3} |
| k5                      | -3.933 \times 10^{-4} |

Figure 7. Chessboard calibration board distortion correction.

Figure 8. Before and after distortion correction.

Figure 9. Actual visible field measurement.

Figure 9, the field of view size of which is 474 mm × 353 mm. The scale of the image size to actual size is 1:0.231, which means the actual size of a single pixel is 0.231 mm.

3) MOTION PARAMETERS AND SAMPLING FREQUENCY
The hardware system was installed on the inspection vehicle at the bottom of the beam of Jiangyin bridge with a movement speed of approximately 0.36 km/h. The inspection vehicle moved the vision system to scan the beam bottom. According to the accuracy of acquisition and the moving speed of the bridge inspection vehicle, the overlapping rate between adjacent images was set to 50%. Therefore, the sampling frequency of the camera was set to 0.5 FPS; that is, one image was collected every 2 seconds.

4) PANORAMA GENERATION
In practice, through the scanning of two scanning strips at the bottom of the box-girder, 17,755 images of the bottom of the steel box girder were obtained in the image acquisition stage. According to the acquisition sequence, every 100 images were stitched and 138 panoramas were obtained for following disease identification. The stitching effect and results are displayed in Figure 10.

C. DEEP NETWORK TRAINING AND EVALUATION
1) DATA SETS
A total of 100 steel box girder images with resolution of 2048 px × 1536 px collected at Jiangyin bridge site were used to form the training data set. Based on the two categories of disease and background, a segmentation template was constructed by manually marking the diseased areas in the images, and some local noise was eliminated using an image processing method. Some of the original images and segmentation templates are shown in Figure 11, where the diseased parts in the original image are marked with yellow. In order to prevent OOM during network training, each original image was divided into 192 images of size 128 × 128; thus, 19,200 images were obtained as the total data set. According to the proportion of 9:0.5:0.5, the total data set was divided into training, validation, and test data (i.e., 17,280 images in the training set, 960 images in the validation set, and 960 images in the test set). Among them, the training set was used for actual data training, the validation set was used to adjust network parameter settings, and the test set was used to evaluate the network segmentation results.
2) MODEL TRAINING

The segmentation network was constructed using the \texttt{keras} deep learning library. The network training process was based on a Windows Server, which was configured with 128 GB memory, an Intel Xeon E5-2630 V4 @ 2.20GHz processor, and an NVIDIA GeForce GTX 1080Ti video card. The training optimizer adopted the Adam algorithm. After training multiple models, a learning rate of 0.001 and a regularization weight of 0.001 were chosen as the best hyper-parameters to set the optimizer. At the same time, a callback function in \texttt{keras} named \texttt{ReduceLROnPlateau} was used. When the network accuracy no longer grew in the verification set for several epochs (set to 10 epochs), the learning rate was automatically scaled down (the scaling coefficient was set to 0.1). In the training process, a training set image and its corresponding segmentation template were transferred into the segmentation network. A total of 16 images were trained at a time. The whole training set image was called an epoch after one training round. The whole training process consisted of 100 epochs. After each epoch, the average loss and learning rate in the current epoch were recorded to determine the best model during processing.

3) MODEL EVALUATION

For a trained model, evaluation of the model is usually carried out to appraise its effects in practice. For binary classification tasks, there will be four cases that the classification results can generate, according to the prediction results of the model and the actual results: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Based on this, the measurement indices accuracy, precision, and recall rate can be defined as:

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},
\]

\[
\text{Precision} = \frac{TP}{TP + FP},
\]

\[
\text{Recall} = \frac{TP}{TP + FN}.
\]

As every position of the output represents the probability of each pixel belonging to the disease category, the PR curve can be obtained by plotting the precision $P$ against recall $R$ in the same co-ordinate system, according to different probability values as the threshold of dividing whether the pixel belongs to which disease. In general, precision and recall are a pair of contradictory performance measures, so it is difficult to have good precision and recall at the same time.

D. COMPARISON BETWEEN DIFFERENT MODELS

In order to achieve the best segmentation effect and explore which architecture is the most suitable for the problem, four sets of models were trained and evaluated; namely, (1) U-Net with VGGNet backbone, (2) U-Net with ResNet backbone, (3) fully convolutional network (FCN) with VGGNet backbone, and (4) FCN with ResNet backbone. The structure of U-Net with VGG-16 backbone is illustrated in Figure 5. Similar to model (1), model (2) has a ResNet backbone in the down-sampling process and a size-corresponding up-sampling process. The only difference between FCN and U-Net with the same backbone is that the latter has merge layer and the former does not. A kernel size of $3^2$ and a stride of 1 were applied to all convolutional and trans-convolutional layers of the models.

1) LOSS, ACCURACY, AND PR CURVES OF BEST MODELS

In this study, four sets of models were evaluated. Loss, accuracy, and precision-recall curves of the best performance
model in each set were plotted, according to the training and validation results. Indices of all models are shown in the same figure, as illustrated in Figures 12, 13, and 14.

The accuracies of all best models reached 90% or higher. However, their mean average precisions (mAPs) varied greatly. In general, the U-Net models performed better than the corresponding FCNs (i.e., normal in the figures). The contrast between U-Net and FCN will be illustrated, in detail, below.

2) COMPARISON BETWEEN DIFFERENT ARCHITECTURES
For each set of models, multiple hyper-parameter combinations were used and various results were obtained. In this section, only the models achieving the best performance in every set were chosen for comparison with the best of the other sets.

a: U-NET BACKBONES
Experiments on U-Net with different backbones were carried out. We chose VGG-11, 13, 16, 19 and ResNet-18, 34, 50, 101 for the down-sampling process, in contrast with the traditional U-Net architecture. The results are shown in Table 3. The VGG-16 backbone can be seen to undoubtedly be the best among all VGG backbones of U-Net, achieving the highest accuracy and mAP. As for ResNet, ResNet-18 obtained both the best accuracy and mAP.

b: CORRESPONDING U-NET AND FCN
As explained before, corresponding U-Net and FCN models had the same structure, but the former had a merge layer while the latter did not. Considering the time cost and the appearance of all models with different backbones, VGG-16 and ResNet-18 were chosen as FCN backbones. The comparison between U-Net and FCN with the same backbone is shown in Table 4.

Table 4 clearly indicates that U-Net (with merge layers) was far more suitable than FCN (without merge layers) for the disease recognition and segmentation problem. No matter what the backbone was, U-Net achieved both better accuracy and mAP. This means that the concatenation between layers of up-sampling and down-sampling remarkably improved the network.

E. QUANTITATIVE EVALUATION OF STEEL BOX GIRDER
In this study, the segmentation threshold was set to be 0.5, in order to ensure 95% accuracy in the recognition results. As seen in Figure 15, TP areas were painted green, which well identified the main parts of the diseases.

By use of the non-overlapping sliding-window method, disease identification was carried out based on the obtained panoramas. The identifications in Figure 10 are displayed in Figure 16: the redder the position in the picture, the higher the possibility of disease.

Furthermore, to carry out a quantitative evaluation of the steel box girder, an index (named disease rate) was defined as:

$$\Phi_d = \frac{\sum p_i}{W \times H},$$

(18)
TABLE 3. Evaluation of different U-Net models.

| backbone    | loss function       | optimizer | epochs | best accuracy | best mAP |
|-------------|---------------------|-----------|--------|---------------|----------|
| VGG-11      | binary cross-entropy| Adam      | 100    | 0.9238        | 0.8932   |
| VGG-13      | binary cross-entropy| Adam      | 100    | 0.9369        | 0.8829   |
| VGG-16      | binary cross-entropy| Adam      | 100    | **0.9548**    | **0.9619**|
| VGG-19      | binary cross-entropy| Adam      | 100    | 0.9412        | 0.9203   |
| ResNet-18   | binary cross-entropy| Adam      | 100    | **0.9492**    | **0.9473**|
| ResNet-34   | binary cross-entropy| Adam      | 100    | 0.9367        | 0.8782   |
| ResNet-50   | binary cross-entropy| Adam      | 100    | 0.9153        | 0.7667   |
| ResNet-101  | binary cross-entropy| Adam      | 100    | 0.9185        | 0.8598   |

TABLE 4. Evaluation results of U-Net and FCN with same backbone.

| backbone    | architecture | loss function       | optimizer | epochs | best accuracy | best mAP |
|-------------|--------------|---------------------|-----------|--------|---------------|----------|
| VGG-16      | U-Net        | binary cross-entropy| Adam      | 100    | **0.9548**    | **0.9619**|
|             | FCN          | binary cross-entropy| Adam      | 100    | 0.8989        | 0.8141   |
| ResNet-18   | U-Net        | binary cross-entropy| Adam      | 100    | **0.9492**    | **0.9473**|
|             | FCN          | binary cross-entropy| Adam      | 100    | 0.8976        | 0.7943   |

FIGURE 15. Original images and corresponding segmentation results.

where $p_i$ indicates the pixels belonging to disease, and $W$ and $H$ represent the width and height of the scanned structural plain, respectively.

Based on this definition, the disease rate of each panorama was calculated. At the same time, the results of each panorama were projected into their corresponding position at the bottom of the box girder, in order to get the distribution map of the health status over the whole bridge, as shown in Figure 17. The red color in the figure represents a higher disease rate, while the blue color represents a lower disease rate. The details of the seven disease rates in the figure are listed in Table 5.

From the statistical results, it is obvious that the disease rate in most areas of the steel box girder bottom were in the range of 2%–4%, and only a few areas were greater than
in this paper, a computer vision-based automated inspection method for the deterioration of the bottom of a steel box girder is proposed. Taking Jiangyin Bridge as the research background, the computer vision system was installed on an inspection vehicle for steel box girder bottom and a panorama generation workflow, including camera calibration, image stitching, and fusion, was used to scan two strips of the girder bottom. The results demonstrated that panoramic image stitching technology based on SIFT feature extraction and matching, bundle adjustment, and multi-end fusion is suitable for panoramic image generation for steel structure surfaces, with only a small probability of failure, as there are many SIFT feature points that can be extracted from areas with diseases, but it is hard to find feature points in areas with good health condition. In addition, for long-sequence image stitching, the bundle adjustment method can effectively reduce the accumulation of errors. The proposed disease identification method, based on a semantic segmentation convolutional neural network, can achieve high-precision pixel-level disease detection and classification. According to our results, the distribution of diseases can be accurately expressed and the health status of the steel box girder bottom of Jiangyin bridge was evaluated by statistical analysis. Except surfaces of steel box girders, our methodology can also be applied on concrete bridge if sufficient data are collected.

Although our methodology showed its reliability in recognizing structural diseases and evaluating the health status of a bridge, there still exist several drawbacks. Firstly, the method is not efficient for acquiring images. Due of the speed limit of the bridge inspection vehicle, its velocity was restricted to 0.36km/h and it took four hours per trip for image acquisition. Secondly, the strips we scanned were only a small part of the steel girder bottom, due to the limitations of camera number and the characteristics of the lenses used; therefore, the full disease distribution on the girder bottom was not obtained. In addition, the lighting condition and unexpected stains on the surface could affect the result of disease identification. Therefore, uniform illumination and a clean surface are required during image acquisition progress.

In future works, several improvements can be carried out: (1) a wide-angle camera can be installed further away from the bridge or a UAV can be applied for the acquisition of data. (2) a new set of data used to distinguish disease of the surface from normal stains can be collected, and the distinguishing ability can also be added into our networks. (3) A real-time monitoring and evaluation system can be developed, which is able to analyze the image immediately after obtaining and which directly outputs a disease distribution map of certain strips. (4) Explorations into more effective network architectures and hyper-parameter combinations can be carried out.

Table 5. Disease image distribution.

| Disease rate | 0%~2% | 2%~4% | 4%~6% | 6%~8% | 8%~10% | 10%~12% | >12% |
|--------------|-------|-------|-------|-------|--------|---------|------|
| Number of panoramas | 24    | 62    | 20    | 16    | 7      | 2       | 1    |
| Proportion   | 17.39%| 44.93%| 18.84%| 11.59%| 5.07%  | 1.45%   | 0.72%|

Figure 17. Disease distribution map of whole bridge.

10%. This indicates that the Jiangyin bridge steel box girder bottom was in good condition, overall. Meanwhile, based on the distributions of the two strips, it is shown that the health status of the box-girder bottom coating near the downstream area was not as good as that near the upstream area, and that the health status on the north side is worse than that on the south side, which well agrees with the real situation: the north side is an industrial area, while the south side is used for living.

VI. CONCLUSION AND FUTURE WORKS

In this paper, a computer vision-based automated inspection method for the deterioration of the bottom of a steel box girder is proposed. Taking Jiangyin Bridge as the research background, the computer vision system was installed on an inspection vehicle for steel box girder bottom and a panorama generation workflow, including camera calibration, image stitching, and fusion, was used to scan two strips of the girder bottom. The results demonstrated that panoramic image stitching technology based on SIFT feature extraction and matching, bundle adjustment, and multi-end fusion is suitable for panoramic image generation for steel structure surfaces, with only a small probability of failure, as there are many SIFT feature points that can be extracted from areas with diseases, but it is hard to find feature points in areas with good health condition. In addition, for long-sequence image stitching, the bundle adjustment method can effectively reduce the accumulation of errors. The proposed disease identification method, based on a semantic segmentation convolutional neural network, can achieve high-precision pixel-level disease detection and classification. According to our results, the distribution of diseases can be accurately expressed and the health status of the steel box girder bottom of Jiangyin bridge was evaluated by statistical analysis. Except surfaces of steel box girders, our methodology can also be applied on concrete bridge if sufficient data are collected.

Although our methodology showed its reliability in recognizing structural diseases and evaluating the health status of a bridge, there still exist several drawbacks. Firstly, the method is not efficient for acquiring images. Due to the speed limit of the bridge inspection vehicle, its velocity was restricted to 0.36km/h and it took four hours per trip for image acquisition. Secondly, the strips we scanned were only a small part of the steel girder bottom, due to the limitations of camera number and the characteristics of the lenses used; therefore, the full disease distribution on the girder bottom was not obtained. In addition, the lighting condition and unexpected stains on the surface could affect the result of disease identification. Therefore, uniform illumination and a clean surface are required during image acquisition progress.

In future works, several improvements can be carried out: (1) a wide-angle camera can be installed further away from the bridge or a UAV can be applied for the acquisition of data. (2) a new set of data used to distinguish disease of the surface from normal stains can be collected, and the distinguishing ability can also be added into our networks. (3) A real-time monitoring and evaluation system can be developed, which is able to analyze the image immediately after obtaining and which directly outputs a disease distribution map of certain strips. (4) Explorations into more effective network architectures and hyper-parameter combinations can be carried out.

REFERENCES

[1] S. Tsao, N. Kehtarnavaz, P. Chan, and R. Lytton, “Image-based expert system approach to distress detection on crc pavement,” J. Transp. Eng., vol. 120, no. 1, pp. 52–64, 1994.
[2] M. R. Jahanshahi and S. F. Masri, “Adaptive vision-based crack detection using 3D scene reconstruction for condition assessment of structures,” Autom. Construction, vol. 22, pp. 567–576, Mar. 2012.
[3] P. Prasanna, K. J. Dana, N. Gacunski, B. B. Basilly, H. M. La, R. S. Lim, and H. Parvardeh, “Automated crack detection on concrete bridges,” IEEE Trans. Autom. Sci. Eng., vol. 13, no. 2, pp. 591–599, Apr. 2016.
[4] G. Li, X. Zhao, K. Fu, R. Fu, and Y. Zhang, “Recognition and evaluation of bridge cracks with modified active contour model and greedy search-based support vector machine,” Autom. Constr., vol. 78, pp. 51–61, Jun. 2017.
[5] Y.-J. Chu, W. Choi, and O. Büyükozkurt, “Deep learning-based crack damage detection using convolutional neural networks,” Comput.-Aided Civil Infrastruct. Eng., vol. 32, no. 5, pp. 361–378, May 2017.
[6] Y. Xu, Y. Bao, J. Chen, W. Zhao, and H. Li, “Surface fatigue crack identification in steel box girder of bridges by a deep fusion convolutional neural network based on consumer-grade camera images,” Struct. Health Monitor., vol. 18, no. 3, pp. 653–674, May 2019.
[7] S. Dorafshan, R. J. Thomas, and M. Maguire, “Comparison of deep convolutional neural networks and edge detectors for image-based crack detection in concrete,” Constr. Building Mater., vol. 186, pp. 1031–1045, Oct. 2018.
[8] D. J. Atta and M. R. Jahanshahi, “Evaluation of deep learning approaches based on convolutional neural networks for corrosion detection,” Struct. Health Monitor., vol. 17, no. 5, pp. 1110–1128, Sep. 2018.
Yiquan Zhang was born in Dalian, China, in 1993. He received the B.S. degree in civil engineering from Tongji University, in 2015, where he is currently pursuing the Ph.D. degree. His research interest is in the application of machine learning, especially in topology optimization, and infrastructural monitoring and maintenance.

Yue Pan received the B.S. degree in geographic information systems from Tongji University, Shanghai, China, in 2011, where he is currently pursuing the Ph.D. degree in civil engineering. He is a Lecturer with the College of Electronics and Information Engineering, Tongji University, and is a Scientific Assistant of the Research Group of Computer Vision and Artificial Intelligence Application in Bridge Engineering. His research interests are in image processing, 3D reconstruction, the Internet of Things, and deep learning.

Bo Peng was born in Harbin, China, in 1994. He received the B.S. and the M.S. degrees in civil engineering from Tongji University, Shanghai, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree in civil engineering with New York University Abu Dhabi, Abu Dhabi, UAE. His current research interests include the applications of machine learning, computer vision, and robotics in civil engineering, especially in the monitoring and maintenance of infrastructure.

Haoran Liu was born in Yichang, China. He is currently pursuing the degree with the School of Civil Engineering, Tongji University, Shanghai. He has a good performance in school and has participated in multiple contests. Having taken good command of major courses, he is interested in calculus and computer science, and is proficient in using Python, MATLAB, and others. He is actively involved in research projects on deep learning and structural disease recognition. In 2018, he won a Second Prize from the National College Mathematics Competition.

Rujin Ma received the Ph.D. degree in bridge and tunnel engineering from Tongji University, China, in 2004. He is an Associate Professor with the Department of Bridge Engineering, Tongji University. He has taken charge of and participated in many national science foundation projects and projects of national science and technology supporting plans. His research interests include the life cycle design theory of bridge, bridge risk assessment, and bridge performance under extreme events, and the maintenance and management of bridge structures.