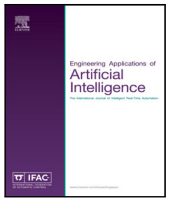
Engineering Applications of Artificial Intelligence 133 (2024) 108574

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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



Research paper

High-resolution cross-scale transformer: A deep learning model for bolt loosening detection based on monocular vision measurement

Wu Tianyi [a], Shang Ke [b], Dai Wei [a], Wang Min [a], Liu Rui [a], Zhou Junxian [a], Liu Jun [a,*]

[a] Department of Mechanical Engineering, City University of Hong Kong, Hong Kong, 999077, China
[b] Department of Mechanical Engineering, Beijing Institute of Technology, Beijing, 100081, China

ARTICLE INFO

Keywords:
Connection loosening detection
High-resolution architecture
Vision transformer
Monocular vision measurement

ABSTRACT

The reliability of bolt connections significantly impacts the operational state and lifespan of industrial equipment. Vision-based noncontact methods exhibit high efficiency in bolt loosening detection. However, limited image features hinder measurement accuracy. To improve bolt loosening detection performance, this paper proposes a novel deep learning backbone, the high-resolution cross-scale transformer, to extract high precision keypoints for bolt three-dimensional model construction. Simultaneously, a monocular vision measurement model is established to get the bolt exposed length and evaluate the connection loosening state. The proposed backbone hybridizes the advantages of high-resolution architecture and transformer, realizing global information aggregation and fine-grained image details. A simplified module, dual-scale multi-head self-attention, is designed to reduce the computational redundancy caused by the implementation of high-resolution multi-branch architecture. In the experiment section, the high-resolution cross-scale transformer outperforms other keypoint detection baselines, achieving the top one performance with 91.6 average precision and 84.9 average recall. The monocular vision measurement model realizes a 0.053 mm error with a 0.028 mm standard deviation, satisfying the industrial implementation requirement. Additionally, the model is tested on different industrial situations and an additional outside dataset, indicating the model's robustness and actual environment adaptability.

1. Introduction

Bolted joints constitute an integral component in various industries, including civil, mechanical, and aerospace engineering, owing to their cost-effectiveness and widespread applicability. However, the harsh operational conditions characterized by excessive load, residual stress, chemical corrosion, and unacceptable assembly (Wang and Song, 2019; Hosseinpour et al., 2023; Mushtaq et al., 2023) often contribute to the occurrence of loosening connections, resulting in engineering accidents and casualties. Thus, it is essential to monitor the health statement of the bolt connection statement to prevent industrial mishaps and elongate the equipment's lifetime.

Currently, there are two categories of bolt connection monitor strategies: contact method and noncontact method. The contact method involves attached sensors, such as ultrasonic sensors (Wang, 2023; Zhang et al., 2019; Hei et al., 2020; Wang et al., 2020a), strain gauges (Ren et al., 2018; Wang et al., 2019b; Duan et al., 2019), and fiber Bragg grating sensors (Miao et al., 2020; Wang et al., 2020c), to monitor connection state. In the measurement process, an input signal produced by the signal generator traverses the bolt connection structure and is captured by attached sensors. Analyzing signal variations enables the computation of the blot connection's preload, offering valuable insights into its tightness condition. However, the attachment of the sensors will impact the structure and thread of target bolt connections. Significantly, the implementation of strain gauges sensors necessitates drilling holes in the bolt, compromising the connection's integrity and shortening the equipment's lifespan. Additionally, the assembly of sensors and their power supply escalate labor costs and the workspace utilization. Furthermore, the sensors are highly susceptible to environmental factors like temperature, humidity, and device vibrations, impacting their reliability.

The noncontact method uses image acquisition devices, such as digital cameras, to get bolt connection figures, which are analyzed with computer vision methods to assess the fasten statement (Wei et al., 2023). Unlike the contact method, the noncontact method avoids attaching or inserting extra components into the bolt connection, guaranteeing the bolt working statement and equipment life. Moreover, image acquisition devices are less susceptible to temperature and humidity compared to electrical or acoustic sensors. Since the bolt loosening process occurs gradually in operations, vision sensors are not required to be

* Corresponding author.
E-mail address: Jun.Liu@cityu.edu.hk (J. Liu).

https://doi.org/10.1016/j.engappai.2024.108574
Received 12 December 2023; Received in revised form 5 April 2024; Accepted 4 May 2024
Available online 13 May 2024

on all the time, reducing the interference of temporary sensor malfunctions. Mazzeo et al. (2004) first introduced an automatic visual system to detect bolt connection defects in railways. They trained a neural to extract the bolt locations and applied wavelet transform and principal component analysis to patterns. The trained classifier can detect the absence or presence of the bolt. This work realized the bolt connection monitor with quantity detected objects, but it can only determine if the bolt is missing instead of assessing for looseness in joints. Feng et al. (2013) broadened the class numbers of losing bolt images, setting three ranks for the fasteners with illumination variation. This expanded the binary classification task into a multiclass classification problem. Ali et al. (2021) deployed modified Faster R-CNN with the unmanned aerial vehicle system, and Zhao et al. (2022) added YOLO-V3 to robotic vision. Both two expanded the application scenarios of bolt looseness detection and reduced computational cost. However, the level setting is based on the subjective experience of the engineers, implying that the classification outcomes are merely qualitative analyses and lack precision. To tackle the previous issue, Cha et al. (2016) combined Hough transform with trained support vector machines to realize the evaluation between bolt loosening states rand rotation angles. Based on this idea, Ramana et al. (2019) introduced the Viola–Jones algorithm to reinforce feature extraction results and realize high performances on smartphone images. Moreover, Wang et al. (2019a) first introduced a semi-supervised learning strategy for angle calculation, and Zhao et al. (2019) configured the neural network on the smartphone. Semi-supervised learning reduced data collection and labeling costs (Jiang et al., 2022), increasing the potential for industrial application. Though the rotation angle calculation can quantitatively evaluate the loosening states of bolt connection, the detection needs to compare the current angle and the initial angle at installation, which is difficult to obtain in actual industrial scenarios. To realize quantitative detection, Gong et al. (2022) detected two types of bolt keypoints and calculated the exposed length of bolts with smartphone images. This method extracts the bolt image keypoint feature and quantitatively evaluates the bolt loosening states. However, they directly adopted the Cascaded Pyramid Network (CPN) (Chen et al., 2018b) without customizing the deep learning network. Meanwhile, the measurement model is only applicable to images with the same length and width, which is unrepresentative in industry settings. These limitations resulted in reduced accuracy and limited generalization. Based on previous work, this paper proposed an innovative monocular vision measurement model with high precision and unrestricted input image size. Simultaneously, this work discusses the keypoint detection in-depth and presents a new backbone for better feature extraction performance.

Deep learning has achieved great success in the field of structural health monitoring (Cha et al., 2017, 2018). However, methods for extracting feature points on bolt threads are still needed. Keypoint detection, a downstream task of deep learning, is active in the human pose detection area with its ability to extract local features. With the locating of human anatomical keypoints, such as head, wrist, and elbow, the 2D human pose is constructed. Compared to traditional deep learning networks, the keypoints backbones focus on aggregating inter-level features (Cai et al., 2020) to extract spatial and semantic information. For example, CPN used a head network to fuse different spatial level features, and Feature Pyramid Networks (Lin et al., 2017a) extended the receptive field from 1/32 to 1/4 with a top-down pathway. Residual Steps Network (RSN) enriched information interaction in the intra-level feature map and realized state-of-the-art (SOTA) in the COCO dataset (Lin et al., 2014). HRnet (Wang et al., 2020b) added semantic information to spatial information in low-resolution sub-networks, becoming a paradigmatic architecture for keypoint detection tasks. However, the performance of these networks is limited by finite receptive fields and strong inductive bias with cascaded convolution kernels (Gu et al., 2022). Beyond the traditional convolution, vision transformer (ViT) (Dosovitskiy et al., 2010) boosted the prediction accuracy in the computer vision downstream tasks, including semantic segmentation (Lewis et al., 2023; Zheng et al., 2021), object detection (Jamil and Roy, 2023; Ma et al., 2023), and video understanding (Bertasius et al., 2021; Zhang et al., 2021). The attention mechanism has also demonstrated its superiority and success in industrial applications (Choi and Cha, 2019; Kang and Cha, 2022; Ali and Cha, 2022; Rosso et al., 2023). In this work, we hybridize the advantages of HRnet and ViT to propose a new backbone, high-resolution cross-scale transformer (HRCSTrans). The HRCSTrans achieves intense expressivity with dynamic aggregation based on the self-attention mechanism. It also enhances fine-grained image details with a cross-resolution multi-branch architecture. To prevent the computational explosion caused by transformer and multi-branch fusion, a simplified self-attention module, dual-scale multi-head self-attention, is designed to replace vanilla multi-head self-attention in transformer blocks. The HRCSTrans achieve top performance on the bolt connection image dataset with 91.6 average precision and 84.9 average recall. It boasts a 4.1 average precision gain compared to the keypoint detection SOTA method RSN and 2.4 points gain compared to the modification baseline HRNet. The HRCSTrans effectively extracts keypoints and ensures high-precision measurement of bolt loosening states.

In summary, noncontact bolt loosening detection face two main challenges. First, the evaluation of bolt loosening states is based on the range of exposed length instead of a precise value. Second, previous noncontact bolt loosening methods focused on extracting the bolt image features instead of three-dimensional reconstruction. Although the multi-view vision is useful in high-precision three-dimensional reconstruction, it is unsuitable for bolt loosening detection due to space requirements and is limited by the increased cost of multiple camera matching. This paper addresses these challenges by proposing a new keypoint detection backbone, HRCSTrans, to extract high-precision keypoints with conventional industrial camera images, and a monocular vision model to calculate the exposed length in the 3D real world with seven 2D keypoints. The flowchart of the proposed system is depicted in Fig. 1. Firstly, the image acquisition platform is constructed and collects bolts images. The images are processed to extract the region of interest (ROI), and the camera parameters are calibrated. Secondly, the ROI is fed to the HRCSTrans, and seven keypoints are obtained with the heatmap. Thirdly, the camera calibration parameters are transported to the monocular vision measurement model, establishing the relationship between different coordinates. With the 2D keypoints, 3D models of the bolt's upper and lower surfaces are constructed. Finally, the spatial distance between these two surfaces is determined, enabling the calculation of the bolt's exposed length for a quantitative detection of the loosening state.

The rest of this paper is organized as follows: Section 2 introduces the total architecture of the HRCSTrans and the sub-module of this new backbone. Section 3 describes the monocular vision measurement model based on seven keypoints. Section 4 compares the proposed method with other deep learning baselines and tests the measurement accuracy in different industrial situations. Section 5 draws the conclusions.

2. High-resolution cross-scale transformer model

To measure the bolt exposed length and evaluate the connection loosening state, the geometric feature points of bolt 2D images need to be extracted. The parallel arc edge of the thread presents a significant obstacle, impeding the precise identification of the bolt tail end's geometric characteristics with traditional edge detectors, such as Canny and Sobel. This section presents a novel keypoint detection deep learning backbone, enabling the extraction of seven bolts' keypoints essential for exposed length calculation. To achieve heightened precision keypoint extraction performance, we integrated HRNet, a SOTA method, with the transformer mechanism, developing a new backbone termed HRCSTrans. The HRCSTrans represents the pioneering application of transformer models in the domain of bolt keypoint detection and achieves top performance compared to other baselines.

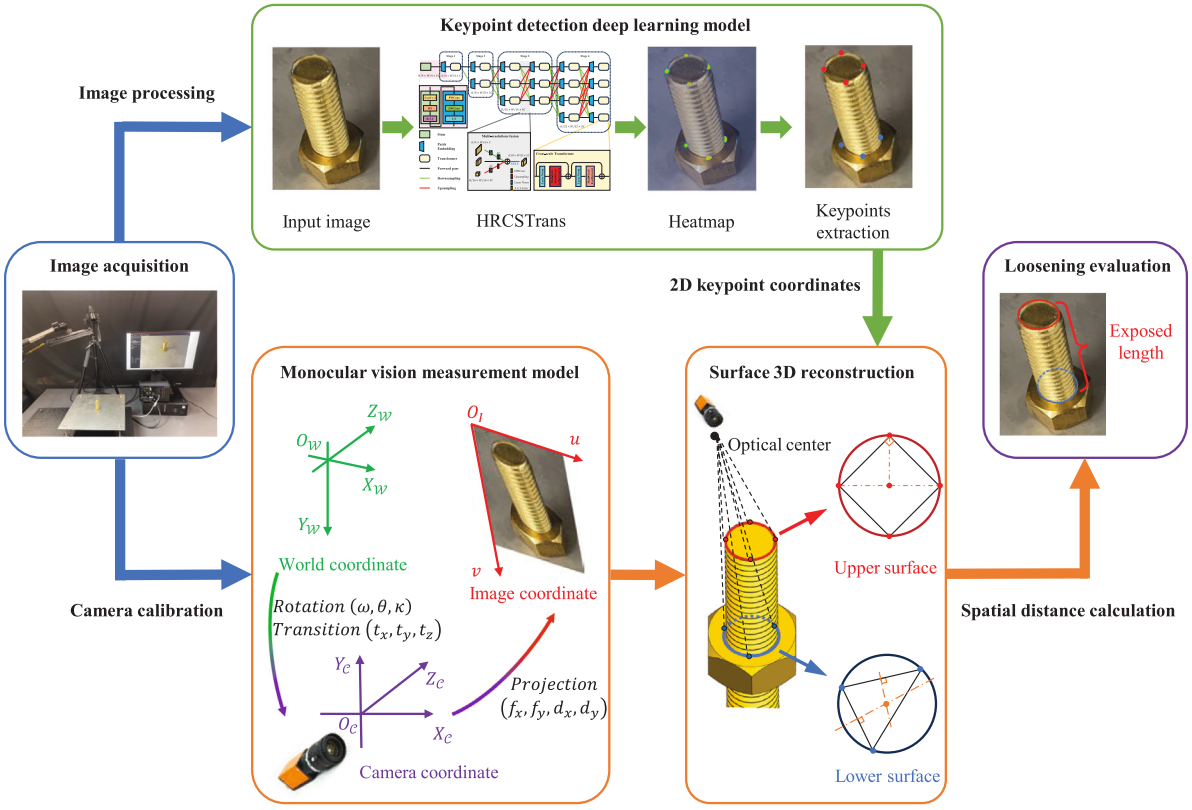


Fig. 1. The vision-based bolt loosening detection method architecture.

2.1. HRCSTrans overview

The overall architecture of HRCSTrans is depicted in Fig. 2. The proposed model consists of a stem head and four-stage feature extractors. The beginning stem is designed to decrease the resolution to 1/4 and reduce the spatial dimension. It contains two stride-2 convolution blocks and realizes downsampling two times, avoiding the explosion in the memory footprint when processing high-resolution images with transformer blocks. The input image is resized to $\frac{H}{4}\times\frac{W}{4}\times 16$ and fed to a Batch Norm and ReLU combination to improve convergence speed. After the convolutional stem, the HRCSTrans deploys feature split in four stages, gradually adding high-to-low resolution streams one by one. The highest resolution feature map is transported in the transformer blocks stream, and resolution streams of the same size are connected in parallel. Different from HRNet, we proposed a new transformer block, cross-scale transformer (CST), to replace the convolution kernel for better intra-level (Cai et al., 2020) information extraction. Before CST blocks in each stream, a patch embedding head is attached to tune the channel size. Due to the multiple branches of HRNet architecture, the number of embedding heads is about three times that of the traditional ViT structure (Dosovitskiy et al., 2010; Liu et al., 2021). To save computational costs, the depth-wise convolution (DWConv) is deployed to replace the convolution block in patch embedding (Chen et al., 2018a). And the point-wise convolution (PWConv) is used to match channels and reinforce patch information interaction for the same reason (Yang et al., 2019). After a layer Norm, the feature map is fed to the CST block to realize intra-level information interaction and keypoint feature extractions. To balance the information processing efficiency and keypoint detection accuracy, we reduce the number of CST blocks and cancel the feature fusion in different resolutions at the first two stages. Stage three and four process fine grained feature maps, critical in keypoint detection tasks. Hence, in these stages, the transformer blocks are repeated, and different features are fused.

After the transformer blocks, a multi-resolution fusion layer is used to fortify the feature map interaction of different resolutions. To guarantee the low-resolution feature maps maintain local position details, all the high-resolution and low-resolution features are fused following the design of HRNet. Unlike the HRNet, HRCSTrans retains the lightweight design in the multi-resolutions fusion layer. The combination of DWConv and PWConv replaces the progressive convolution in the downsampling process. The DWConv shrinks the spatial dimension while the PWConv expands the channel to match the low-resolution feature size. Each convolution is followed by a normalization block to accelerate the learning convergence speed and avoid the problem of vanishing gradients. In the upsampling process, a PWConv is first employed to half the channel length, and a norm layer is followed to normalize the feature. The normalized feature is fed to the nearest upsampling model to upscale the spatial dimension. For the same resolution features, they are directly passed through the forward skip and added with the features from the downsampling and upsampling path. Finally, the fused features are fed to an activation function and transported to the next patch embedding. Inspired by the convolution and transformer hybrid diagram (Liu et al., 2022), the Layer Norm and GELU activation are chosen instead of the universal Batch Norm and ReLU.

2.2. Cross-scale transformer block and dual-scale multi-head self-attention

HRNet constructs a rich receptive field pyramid and has succeeded in the keypoint detection field through information fusion at different resolutions. Nevertheless, HRNet focuses on the information transaction on different size feature maps and ignores the feature extraction in the same resolution (Cai et al., 2020). ViT, famous for its self-attention mechanisms, is good at establishing information interaction in the intra-level feature. To modify the performance of the current backbone in the keypoint detection task, we combine the advantages of HRNet

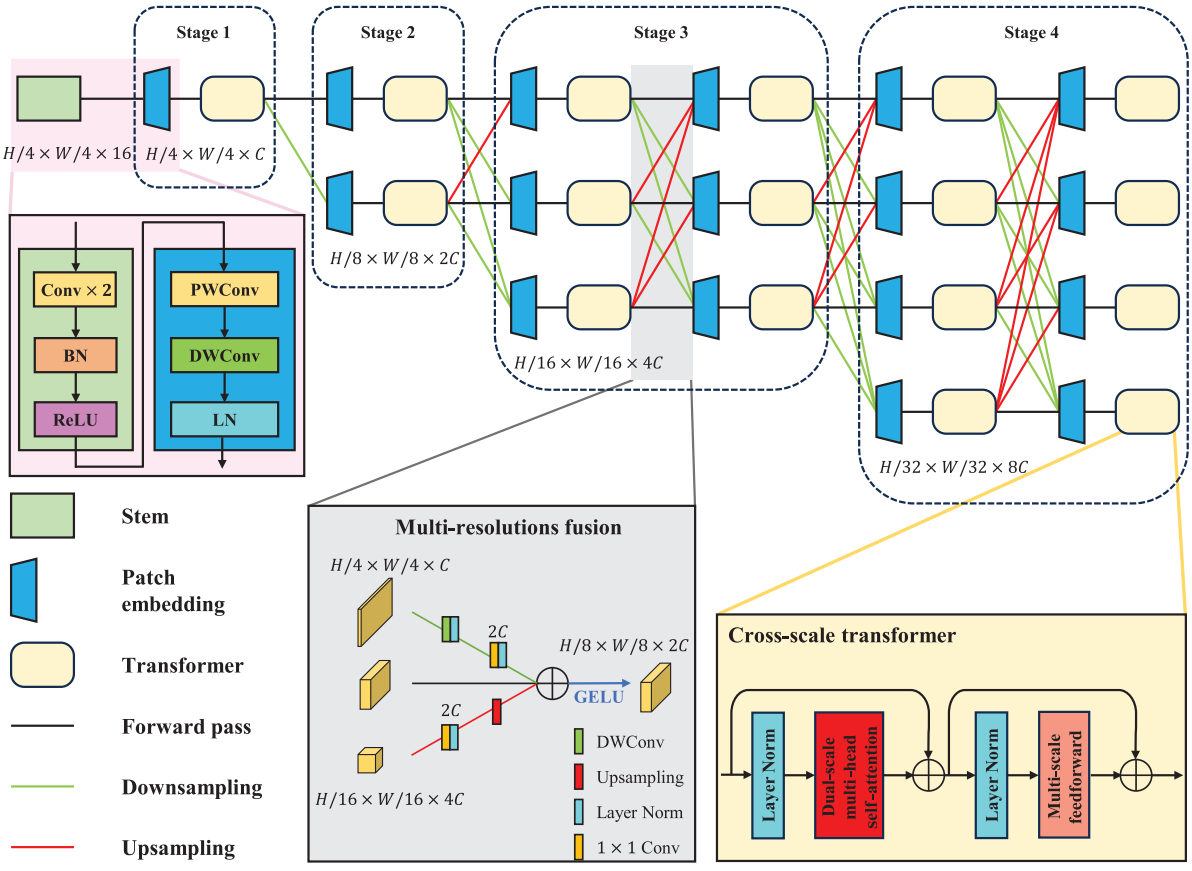


Fig. 2. The overview of high-resolution cross-scale transformer.

and ViT to build our HRCSTrans. Considering the multi-branch structure of HRNet and the high computational cost of self-attention, directly using transformer blocks to replace the convolutions in HRNet will cause parameter explosion. To hybridize the advantages of HRNet and ViT and avoid parameter explosion, we modified the transformer block and proposed a new module, CST, as shown in Fig. 2. The input feature is fed to a Layer Norm and transported to the dual-scale multi-head self-attention (DS-MHSA) module. After skip link connection, the feature is fed to a feedforward network to reinforce channel wise information interaction.

The DS-MHSA module is shown in Fig. 3. Inspired by the CSWin (Dong et al., 2022) and HRVit (Gu et al., 2022), the input feature, denoted as $x \in \mathbb{R}^{H\times W\times C}$, is split to two sides in the channel wise for reduce computational parameters. After splitting, the feature map is transported to a dual branch architecture to pass through the self-attention mechanism in different scales. Different from classic transformer block, the patch embedding in our cross-scale transformer uses a rectangle window instead of a square. The height wise feature $x_h \in \mathbb{R}^{H\times W\times C/2}$ and same size width wise feature x_w are transported different branch for different scale window patching. Compared to vanilla vision transformer, this design interacts feature in the same channel with different window scale, and it also avoids the double computation cost caused by window shuffle and repeated attention block (Dong et al., 2022). In the upper branch, the input feature x_h is patched in height wise and presented as $x_h = \left[x_h^1, x_h^2, \ldots, x_h^i, \ldots, x_h^I\right], x^i \in \mathbb{R}^{H^{win}\times W\times C/2}$. The x_h^i means the i_{th} patch of input feature and the H^{win} is the window size in the height scale. The total numbers of patch I equals to the height size divided by height scale window size, defined as $I = H/H^{win}$. After patching, the feature map is fed to the height scale multi-head self-attention (HS-MHSA) block. At the beginning, the x_h^i is passed to a QKV three branches structure, defined as Eq. (1):

$$y_j^i = Softmax\left(W_j^Q x_h^i \left(W_j^K x_h^i\right)^T / \sqrt{C^j}\right) W_j^V x_h^i \tag{1}$$

where y_j^i is j_{th} head output of x_h^i in the QKV branches and C^j is the channel size of each head feature. W_j^Q, W_j^K, W_j^V corresponds to the parameter metrics of Query, Key and Values, and the metric size equals to $\mathbb{R}^{C^j\times C/2}$. In this work, the parameters of W_j^K and W_j^V are shared to save computational cost and release reduction caused by multi-resolution architecture. The Key metric is transported and multiples with the Query metric. The expanded feature is divided by the square root value of channel length and fed to the softmax function. Finally, the uniformization value multiples with the shared value metric and the output y_j^i is obtained. Unlike vanilla self-attention blocks, the subsidiary skip link is added to the shared value branch to reinforce local feature aggregation. The function of the subsidiary skip link is shown in Eq. (2):

$$z_j^i = y_j^i + DWConv\left(\delta\left(W_j^K x_h^i\right)\right) \tag{2}$$

The $W_j^K x_h^i$ is the shared value metric with input and it is passed through an H-swish activation function, denoted as δ. A DWConv is deployed to enrich local information interactions ignored by the self-attention mechanism. Then, the link output is added to the QKV branches' output y_j^i and get the j_{th} head output z_j^i. Before the window concat, a feature fusion module f is exploited to mix different head features, and it is defined as:

$$f\left(z^i\right) = LN\left(\delta\left(DWConv\left[z_1^i, z_2^i, \ldots, z_j^i, \ldots, z_J^i\right]\right)\right) \tag{3}$$

The z^i presents the feature fusion results corresponding to the x_h^i input. The J is the total head number and equals to $C/\left(2C^j\right)$. All the head outputs z_j^i are concatenated together and passed through a DWConv to inject inductive bias to facilitate training. An H-swish activation function and layer norm are followed to improve convergence speed. Simultaneous Eqs. (1)–(3) and concatenating all the window features, the output of HS-MHSA with total height scale patching input

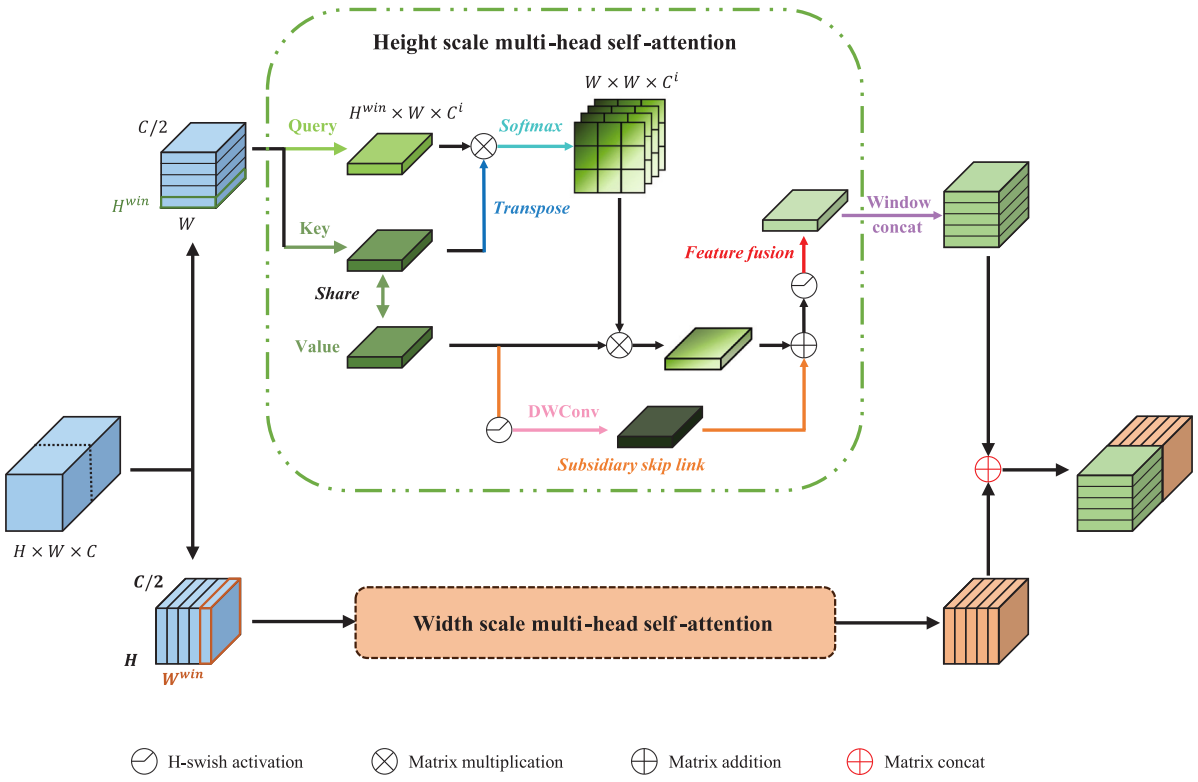


Fig. 3. Architecture of the dual-scale multi-head self-attention module.

x_h is calculated. The width scale multi-head self-attention (WS-MHSA) composition is similar to the HS-MHSA, including QKV branches, a subsidiary skip link, and a feature fusion layer. So, the details of WS-MHSA are omitted in Fig. 3. Different with HS-MHSA, the patching splitting of WS-MHSA is along the width scale instead of the height scale. It guarantees that the feature in the same channel can interact with the feature in different patches, and the various global information will effectively improve the keypoint extraction effect. In the end, the output of HS-MHSA and WS-MHSA are concatenated in channel wise and it is the final output of DS-MHSA block.

2.3. Multi-scale feedforward

In the previous section, the DS-MHSA block realizes a cross-scale global information aggregation with limited parameters. Meanwhile, the subsidiary skip link guarantees the preservation of detailed information. However, the two split features are spliced together directly after passing through the dual attention architecture. It causes a lack of information fusion between these two parts, leading to severe information isolation. To solve this, in the CST module (Fig. 2), a multi-scale feedforward (MSF) block is proposed to integrate the feature map processed by the DS-MHSA block. The details of MSF are shown in Fig. 4. The input feature $\mathbb{R}^{H\times W\times C}$ is expanded k times in the channel wise, and this is the first fusion of the segmented feature transmitted by the DS-MHSA block's dual branch architecture. Followed by the PWConv expansion, the feature map is split into three parts and transported to three branches. To compensate for the lack of local information interaction caused by the transformer window patching, the MSF block deploys three different scales DWConvs. Different from the design of Segformer (Xie et al., 2021), the expansion ratio k is controlled as three to respond to the branches number, and the 3×3, 5×5 and 7×7 kernel sizes are implemented on different branches to enrich perspective fields. After different scales convolution, the feature map is hybrid by the matrix concat. Passing through a GELU activation function, the feature map gets the second time channel wise

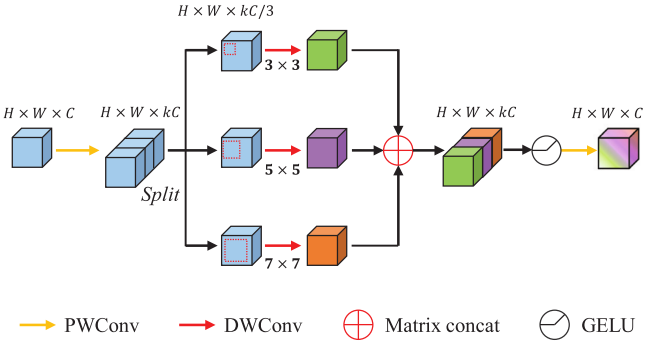


Fig. 4. Architecture of the multi-scale feedforward block.

integration with a PWConv, and the size is reverted to $\mathbb{R}^{H\times W\times C}$. The MSF block integrates the channel wise features divided by the HS-MHSA and WS-MHSA. It also achieves multiple receptive expansions using a three-branch architecture. With the aid of the MSF block, the HRSCTrans accomplishes high precision keypoints extraction, which positively impacts the bolt 3D construction and loosening detection.

3. Monocular vision measurement model

Employing our HRCSTrans model, the precision coordinates of seven keypoints are derived from the 2D bolt image. Nonetheless, the construction of a 3D bolt model, particularly in a monocular vision system, continues to pose a significant challenge. In light of this, we present a monocular vision measurement model for 3D bolt reconstruction utilizing limited keypoints. Seven keypoints are used in this model, while four keypoints are used to determine the space coordinates of bolt's upper surface and the other three keypoints are used to determine the lower surface. The exposed length is calculated by the spatial distance between these two surfaces, implemented to evaluate the loosening state.

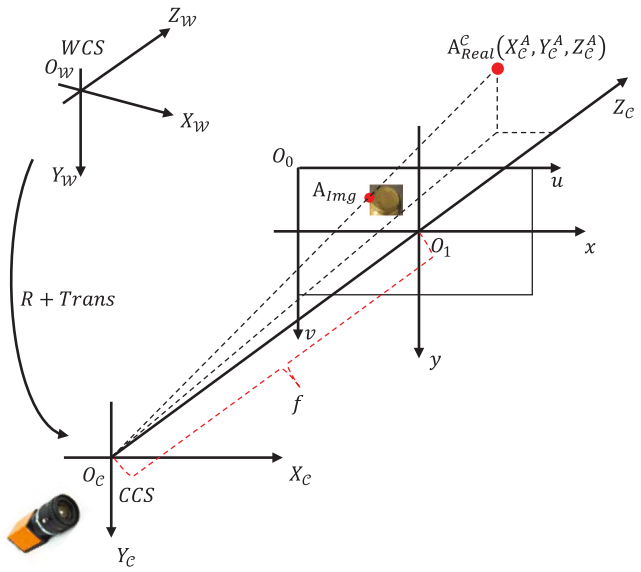


Fig. 5. Keypoint 3D coordinate mapping.

3.1. The image keypoint mapping with 3D space

The camera system operates on the principle of small aperture imaging, and the keypoint 3D coordinate mapping relationship is constructed as shown in Fig. 5. This mapping process involves four coordinate systems, including the world coordinate system (*WCS*), the camera coordinate system (*CCS*), the image coordinate system and the pixel coordinate system. The *WCS* represents the physical location of the object in the real world. Simultaneously, the *CCS* is established with the camera's optical center (O_C) serving as its coordinate origin. The image coordinate system (O_1xy) and the pixel coordinate system correspond to the 2D image plane. The former represents the physical scale, while the latter represents the pixel scale. The coordinate (x, y) of O_1xy and the coordinate (u, v) is related to image size d_x and d_y, i.e., $u = x/d_x$. The line O_CO_1, connecting the origin of *CCS* and O_1xy, is perpendicular to the image plane, and the length of O_CO_1 is the focal length f. The $A_{Img}\left(u^A, v^A\right)$ is the pixel coordinate of the extracted keypoint A and A_{Real} represents the 3D position of A in the real world. The A_{Real} has two coordinate formats: the $A^C_{Real}\left(X^A_C, Y^A_C, Z^A_C\right)$ under *CCS* and the $A^{\mathcal{W}}_{Real}\left(X^A_{\mathcal{W}}, Y^A_{\mathcal{W}}, Z^A_{\mathcal{W}}\right)$ under *WCS*. To obtain the bolt 3D model and the exposed length in the real world, the mapping relation of A_{Img} and $A^{\mathcal{W}}_{Real}$ is constructed as Eq. (4).

$$Z^A_C\begin{bmatrix}u^A\\ v^A\\ 1\end{bmatrix} = \boldsymbol{I}\begin{bmatrix}\boldsymbol{R} & \boldsymbol{Trans}\\ 0 & 1\end{bmatrix}\begin{bmatrix}X^A_{\mathcal{W}}\\ Y^A_{\mathcal{W}}\\ Z^A_{\mathcal{W}}\\ 1\end{bmatrix} \tag{4}$$

The matrix $\boldsymbol{I} \in \mathbb{R}^{3\times4}$ presents the camera intrinsic parameters, and $\boldsymbol{R} \in \mathbb{R}^{3\times3}$, $\boldsymbol{Trans} \in \mathbb{R}^{1\times3}$ present the camera extrinsic parameters. $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ is rotation transformation from *WCS* to *CCS*, and $\boldsymbol{Trans} \in \mathbb{R}^{1\times3}$ corresponds to the translation transformation. All the camera parameters are calibrated by the flexible camera calibration algorithm (Zhang, 1999), and the calibration results are shown in Section 4.1, Table 1. The details of transformation are defined as Eqs. (5)–(7).

$$\boldsymbol{I} = \begin{bmatrix}f/d_x & 0 & u_0 & 0\\ 0 & f/d_y & v_0 & 0\\ 0 & 0 & 1 & 0\end{bmatrix} \tag{5}$$

$$\boldsymbol{R}(\omega,\theta,\kappa) = \begin{bmatrix}1 & 0 & 0\\ 0 & \cos\omega & -\sin\omega\\ 0 & \sin\omega & \cos\omega\end{bmatrix}\boldsymbol{R'}$$
$$\boldsymbol{R'} = \begin{bmatrix}\cos\theta & 0 & \sin\theta\\ 0 & 1 & 0\\ -\sin\theta & 0 & \cos\theta\end{bmatrix}\begin{bmatrix}\cos\kappa & -\sin\kappa & 0\\ \sin\kappa & \cos\kappa & 0\\ 0 & 0 & 1\end{bmatrix} \tag{6}$$

$$\boldsymbol{Trans} = \begin{bmatrix}t_x & t_y & t_z\end{bmatrix}^T \tag{7}$$

The u_0 and v_0 corresponds to the distance between O_0 and O_1 on u axis and v axis. ω, θ and κ mean the rotation angle along $X_{\mathcal{W}}$ axis, $Y_{\mathcal{W}}$ axis and $Z_{\mathcal{W}}$ axis, while t_x, t_y and t_z are the translation along these directions. For convenience of calculations, we combine Eqs. (5)–(7) and simplify as below:

$$Z^A_C\begin{bmatrix}u^A\\ v^A\\ 1\end{bmatrix} = \boldsymbol{M}\begin{bmatrix}X^A_{\mathcal{W}}\\ Y^A_{\mathcal{W}}\\ Z^A_{\mathcal{W}}\\ 1\end{bmatrix} \tag{8}$$

The matrix $\boldsymbol{M} \in \mathbb{R}^{3\times4}$ is the combination of camera intrinsic parameters and extrinsic parameters. Based on Section 2, the pixel coordinate $A_{Img}\left(u^A, v^A\right)$ of keypoint A is obtained, and Eq. (8) converts into a four-quad system with four unknowns, $X^A_{\mathcal{W}}$, $Y^A_{\mathcal{W}}$, $Z^A_{\mathcal{W}}$ and Z^A_C. By substitution, the coordinates of Z^A_C is eliminated and Eq. (8) is rewritten as:

$$\begin{bmatrix}u_Am_{31}-m_{11} & v_Am_{31}-m_{21}\\ u_Am_{31}-m_{11} & v_Am_{31}-m_{21}\\ u_Am_{33}-m_{13} & v_Am_{31}-m_{21}\end{bmatrix}^T\begin{bmatrix}X^A_{\mathcal{W}}\\ Y^A_{\mathcal{W}}\\ Z^A_{\mathcal{W}}\end{bmatrix} = \begin{bmatrix}m_{14}-u_Am_{34}\\ m_{24}-v_Am_{34}\end{bmatrix} \tag{9}$$

The $m_{i,j}$, $i \in [1..3], j \in [1..4]$ means the element in i_{th} row and j_{th} column of matrix $\boldsymbol{M}$. Eq. (9) is a ternary system of two equations representing the space line O_CA_{real} in *WCS*. In the same way, based on the other six keypoints coordinate $B_{Img}\left(u^B, v^B\right), C_{Img}\left(u^C, v^C\right), \ldots, G_{Img}\left(u^G, v^G\right)$, the connection line between the camera light center and the keypoint 3D real position is calculated, defined as $O_CB_{real}, O_CC_{real}, \ldots, O_CG_{real}$.

3.2. Spatial distance calculations

Building upon the mapping relationship between keypoint image coordinates and 3D real positions in *WCS*, the calculation of the connection line function between the camera optical center and the keypoint 3D real position is performed. Although these mapping relationships facilitate the transformation from the 2D image to 3D space, achieving precise 3D coordinates remains elusive. To attain the bolt 3D information, the position of the 3D keypoint in the connection line must be affirmed. This section introduces a spatial distance calculation model, incorporating bolt voxel information, designed to determine the actual 3D position of the keypoint coordinate and the exposed bolt's length.

In the spatial distance calculation process, the seven keypoints construct two surfaces, and the distance between the two surfaces corresponds to the exposed length. For the upper face, determined by keypoints A, B, C and D, the keypoints are located at the axis endpoints of the 2D ellipse. In the 3D mapping space, the keypoints' *WCS* coordinates quarter the circumference of the upper surface circle. The calculation process is shown in Fig. 6. based on the bolt voxel information. Similar to the model in Section 3.1, the point O_C means the camera's optical center and $A_{real}, \ldots, D_{real}$ are the*WCS* coordinates of four upper surface keypoints. The H^1_{real} is the center of the circle $A_{real}B_{real}C_{real}$. Based on the bolt voxel information, the space line and bolt radius are related by Eq. (10).

$$\left|A_{real}B_{real}\right| = \left|B_{real}C_{real}\right| = \sqrt{2}r \tag{10}$$

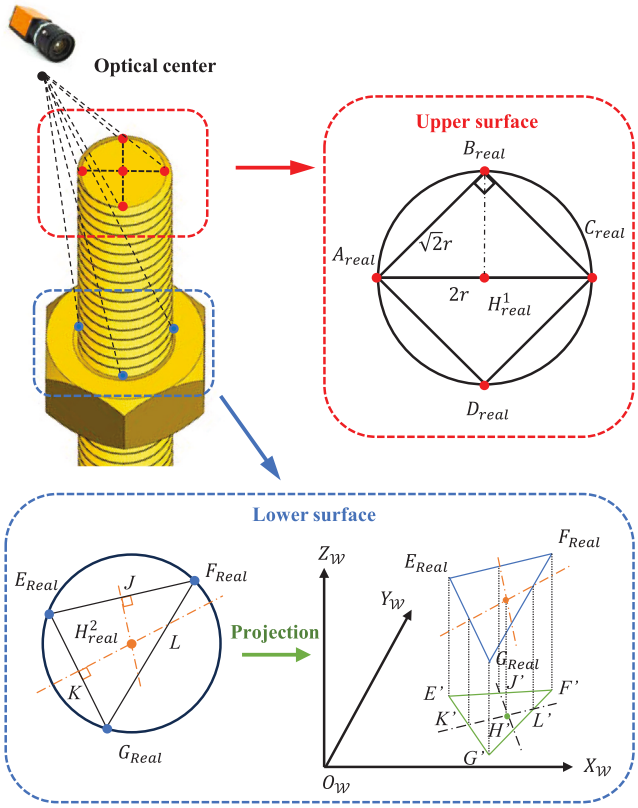


Fig. 6. Surfaces 3D reconstruction.

Table 1
Calibration parameters of the camera.

Parameters	Type	Value
f_x (mm)	Intrinsic parameters	8.49
f_y (mm)		8.49
d_x (μm)		4.65
d_y (μm)		4.65
ω	Rotation angle	2.856
θ		3.056
κ		1.240
t_x	Transition distance	0.199
t_y		−0.085
t_z		1.984

where $|A_{real}B_{real}|$ is the length of the line $A_{real}B_{real}$ and r means the radius of the target bolt. Introduce the pixel coordinates $B_{Img}\left(u^B, v^B\right)$ and $C_{Img}\left(u^C, v^C\right)$ to Eq. (9) and the actual radius to Eq. (10), the *WCS* coordinate of A_{real}, B_{real} and C_{real} can be calculated. In order to improve the accuracy of the calculation result and solve the multiple-solution problems that may occur in spatial symmetry, we use the triangle $B_{real}C_{real}D_{real}$, $A_{real}C_{real}D_{real}$ and $A_{real}B_{real}D_{real}$ to repeat the Eq. (10) calculation process and calculate the average of all keypoints *WCS* coordinates as the final results. Based on this process, the four keypoints A_{real}, B_{real}, C_{real}, D_{real} and center $H^1_{real}\left(X^{H^1}_W, Y^{H^1}_W, Z^{H^1}_W\right)$ is obtained. The surface $A_{real}B_{real}C_{real}$ is also calculated, defined as:

$$\begin{vmatrix} x - X^A_W & y - Y^A_W & z - Z^A_W \\ X^B_W - X^A_W & Y^B_W - Y^A_W & Z^B_W - Z^A_W \\ X^C_W - X^B_W & Y^C_W - Y^B_W & Z^C_W - Z^B_W \end{vmatrix} = 0 \tag{11}$$

For the lower surface, the keypoints cannot be positioned on the axis of the 2D ellipse due to occlusion. So, the angle information of the equilateral triangle cannot be used. To calculate the surface $E_{real}F_{real}G_{real}$, the 3D keypoints are projected to a 2D surface, as shown in Fig. 6. The E_{real}, F_{real} and G_{real} are the lower surface keypoints in *WCS*. Due to the triangle $E_{real}F_{real}G_{real}$ has a circumcircle with r radius, the perpendicular line of the string is used to determine the center of the circle. The point J is the intersection point of string $E_{real}F_{real}$ and its perpendicular bisector, while point K is the intersection point of the string $E_{real}G_{real}$ and its perpendicular bisector. These two perpendicular bisectors intersect at point H^2_{real}, the center of circle $E_{real}F_{real}G_{real}$. Additionally, line KH^2_{real} interaction line $F_{real}G_{real}$ at point L. Different from equilateral triangles, the included angle of triangle $E_{real}F_{real}G_{real}$ is unknown. To solve this, triangle $E_{real}F_{real}G_{real}$ is projected to plane $O_wX_wY_w$, to convert a 3D space problem into a 2D plane solution, as shown in Fig. 6. The $E'\left(X^E_W, Y^E_W\right)$ is the projection of E_{real} and the coordinates of other projection points are also the same. The function of line $J'H'$ and $K'L'$ is calculated as below:

$$\begin{cases} \left(x - X^J_W\right)\left(Y^H_W - Y^J_W\right) = \left(y - Y^J_W\right)\left(X^H_W - X^J_W\right) \\ \left(x - X^J_W\right)\left(Y^H_W - Y^J_W\right) = \left(y - Y^J_W\right)\left(X^H_W - X^J_W\right) \end{cases} \tag{12}$$

Based on Eq. (9), the coordination of $H'\left(X^H_W, Y^H_W\right)$, corresponding to the X_w and Y_w coordination of H^2_{real}. Similar to this, the coordinate of $L'\left(X^L_W, Y^L_W\right)$ is calculated by interaction line $K'H'$ and $G'F'$ as shown below:

$$\begin{cases} \left(x - X^K_W\right)\left(Y^H_W - Y^K_W\right) = \left(y - Y^K_W\right)\left(X^H_W - X^K_W\right) \\ \left(x - X^G_W\right)\left(Y^F_W - Y^G_W\right) = \left(y - Y^G_W\right)\left(X^F_W - X^G_W\right) \end{cases} \tag{13}$$

To calculate the Z_w coordinate of H^2_{real}, the ratio between triangle KLL'' and $KL''H^2_{real}$ is used as Eq. (14):

$$\frac{Z^{H^2}_W - Z^K_W}{Z^L_W - Z^K_W} = \frac{|KH''|}{|KL''|} = \frac{|K'H'|}{|K'L'|} \tag{14}$$

Based on Eqs. (4) and (9), the point of circle $E_{real}F_{real}G_{real}$ is presented by 2D image coordinates. The circumcircle of circle $E_{real}F_{real}G_{real}$ is used to construct the additional constraint relationship between the *WCS* coordinates and the real bolt vox:

$$\left(X^i_W - X^{H^2}_W\right)^2 + \left(Y^i_W - Y^{H^2}_W\right)^2 + \left(Z^i_W - Z^{H^2}_W\right)^2 = r^2 \tag{15}$$

where $i \in \{E, F, G\}$ indicates the point name in the circle and r corresponds to the bolt radius. To solve the multiple solution question, the boundary regulation surface $A_{real}B_{real}C_{real}D_{real}$ parallel to surface $E_{real}F_{real}G_{real}$. Simultaneous Eqs. (4) and (11)–(15), the keypoints *WCS* coordinates are calculated. With the keypoint coordinates of the upper surface, coordinates of point H^1_{real} and H^2_{real} can be obtained. Finally, the space line length $\left|H^1_{real}H^2_{real}\right|$ is calculated, corresponding to the bolt exposed length. The loosening state of the bolt connection can be evaluated by the exposed length.

4. Experiment

To validate the accuracy and efficacy of our method, we implemented an image acquisition platform grounded in the monocular vision measurement model. A bolt image dataset is built via the platform and tested on the HRCSTrans backbone. The details of the experiment process are demonstrated in this section.

4.1. Image acquisition preparation

As depicted in Fig. 7, a specialized image acquisition platform is devised to ensure image quality and ground truth accuracy. The chosen camera is the HIKvision industrial camera MV-CA050-10GM (2448 × 2048 camera resolution) coupled with MVL-MF0828M Len, which is connected to a PC to store the bolt images. Illumination is provided by the TH2-160X120SW LED light source, supported by

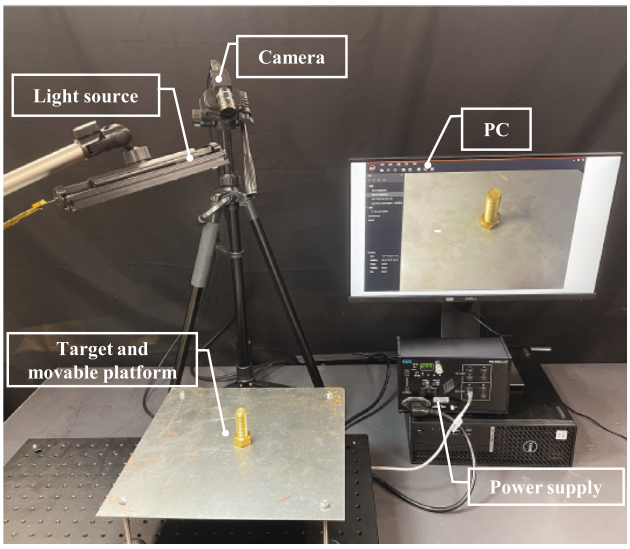


Fig. 7. Image acquisition platform.

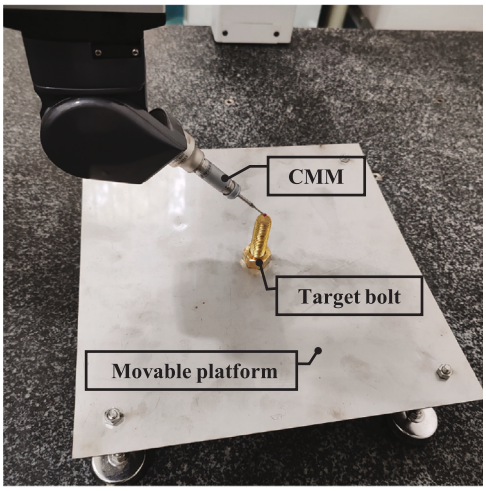


Fig. 8. Ground truth establishment.

the CCS.PD3-5024-4-PI light power supply. A mobile platform is used to assemble the target bolt, ensuring the accuracy of measurement ground truth. The bolts are fastened by a torque wrench with 10 Nm torque. Following image acquisition, the mobile platform is moved to the coordinate measuring machine (CMM) to get the exposed length, as shown in Fig. 8. In this paper, the CMM is the Hexagon Leitz PMM-XI12107 CMM with $\pm(0.5 + L/500)$ μm precision, enhancing the reliability of exposed length measurement.

Before identifying the keypoint *WCS* coordinates, the industrial camera is calibrated by flexible camera calibration (Zhang, 1999), implemented with OpenCV and Python (3.8.11). The calibration results are shown in Table 1. The intrinsic parameters are used in Eq. (5), while the rotation angle and transition distance constitute the extrinsic parameters, feeding to Eqs. (5) and (6). Camera distortion elimination is omitted in the calibration because industrial camera images can realize high-precision bolt loosening detection. Image augmentation and distortion elimination are needed when applying to complex data distribution (Hong et al., 2018).

Table 2
Training hyperparameter.

Config	Value
Dataset size	600 training, 150 testing
Camera resolution	2448 × 2048
Input image size	384 × 288
Batch size	16
Epoch	1000
Drop rate	0.2
Learning rate	0.002
Optimizer	Adam, betas = (0.9,0.999)
StepLR	Step size = 20, gamma = 0.8
Decay rate	0.0001

4.2. Keypoint detection result

A bolt 2D image dataset, coupled with 3D ground truth, is established based on the image acquisition platform. The M14 bolt and nut, featuring a two-millimeter pitch and composed of brass, are employed in this study, encompassing an exposed length range from 10 mm to 60 mm. Keypoints are annotated by LabelMe and converted into COCO (Lin et al., 2014) data format for the deep learning process. The labeled dataset comprises 750 images, with 80% (600 images) designated for training and the remaining 20% (150 images) reserved for testing. Addressing the specific input image size requirement for keypoint detection task, the RetinaNet (Lin et al., 2017b) is deployed for the region of interest extraction. The extraneous background is removed, and the input image used for keypoint detection is modified to an aspect ratio of four to three.

After image processing and ground truth measurement, the labeled images are fed to the HRCSTrans. The region of interest extractor and HRCSTrans are programmed in Pytorch (1.9.0) with Python (3.8.11). The model is trained on a workstation with a CPU of Intel Xeon Platinum 8375C @2.90 GHz and an NVIDIA GeForce RTX 3090 GPU with 24 GB memory using the PyCharm. In the training process, the batch size of 16 is selected with the 384 × 288 input image size, and the Adam is employed as the optimizer with a 0.002 learning rate and $0.9\beta_1$, $0.999\beta_2$. To guarantee the training speed and model convergence accuracy, the learning rate decay is introduced in the training strategy. The learning rate declines in 0.0001 rate and renews in every 20 epochs with 0.8 γ. The training process lasts 1000 epochs with a 0.2 drop rate, and specific values of the hyperparameters are shown in Table 2.

The training process is illustrated in Fig. 9. It is evident that the training loss and training accuracy undergo rapid changes and maintain the initial trend in the first 50 epochs, attributed to the significant learning rate at the beginning of training. With the learning rate decay training strategy, the training loss gradually stabilizes, and the training accuracy peaks, indicating successful model convergence. Compared to the HRnet, the HRCSTrans has a higher initial loss (0.1687 compared to 0.0988) and converges more slowly in the first 50 epochs, owing to the additional transformer blocks. After the 50th epoch, the converge speed of HRCSTrans gradually exceeds that of HRnet. Both models achieve stable training loss by the 900th epoch. Ultimately, the HRCSTrans exhibits a lower final training loss (0.0002 compared to 0.0037), underscoring the superior training performance.

To illustrate the effectiveness of our method, we compare HRCSTrans with other SOTA keypoint backbones, including HRnet, RSN, SimpleBase, and CPN. Additionally, we contrast our method with the bottom-up paradigm, the OpenPose. The standard evaluation metric employed is Object Keypoint Similarity (OKS), defined as Eq. (16).

$$\mathrm{OKS} = \frac{\sum_i exp\left(-d_i^2/2s^2k_i^2\right)\delta\left(v_i > 0\right)}{\sum_i \delta\left(v_i > 0\right)} \tag{16}$$

where d_i is the Euclidean distance between the prediction keypoint image coordinates and the corresponding labeled ground truth. v_i corresponds to the visibility flag of the ground truth, s is the object

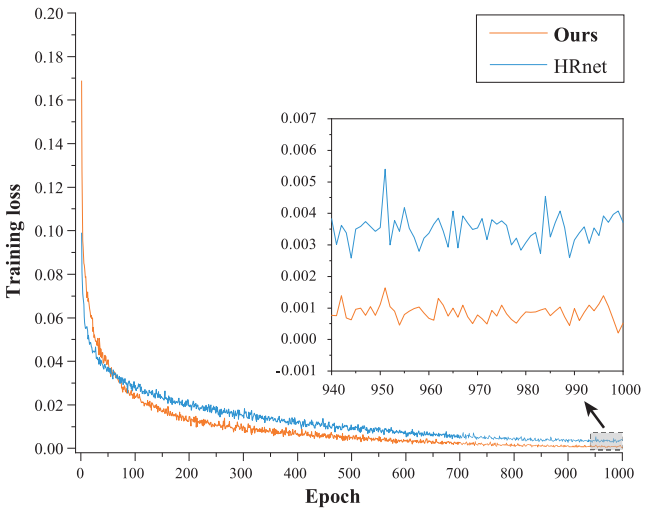


Fig. 9. Training process.

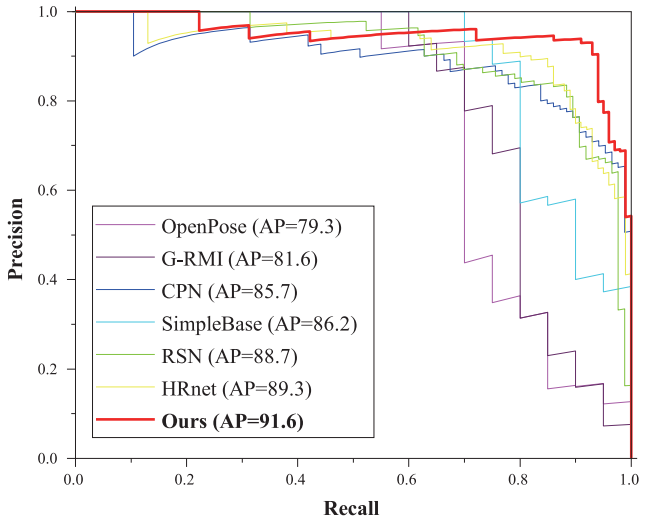


Fig. 10. AP comparison with other methods.

Table 3
Performance comparison with other baselines.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
OpenPose	71.3	79.3	68.7	66.1	69.7	68.5
G-RMI	75.2	81.6	73.9	74.3	73.0	75.4
CPN	81.5	85.7	79.0	79.2	77.5	78.9
SimpleBase	83.4	86.2	81.7	79.3	79.0	79.4
RSN	85.1	88.7	83.6	82.9	81.6	82.3
HRnet	86.8	89.3	82.5	83.5	83.1	81.7
Ours	**89.2**	**91.6**	**87.1**	**86.1**	**87.8**	**84.9**

scale, and k_i is a falloff controlling constant mapping to each keypoints. According to the COCO data format evaluation criteria (Lin et al., 2014), the average precision (AP) based on various OKS and recall scores are utilized to assess keypoint detection performance, as detailed in Table 3. The AP^{50} denotes the AP value at OKS = 0.5, while AP^{75} represents the AP at OKS = 0.75. AP is the mean value of AP scores at ten different OKS positions ($OKS = 0.45 + 0.05\lambda,\ \{\lambda \in [1, 10] \,|\, \lambda = Z\}$). The AP^{M} focuses on testing medium objects, while AP^{L} takes care of large objects. The AR means the average recall scores at ten OKS positions.

The prediction results of the HRCSTrans and other SOTA methods are reported in Table 3. Our method reaches the best AP score, 89.2, and has 2.4 points gain than the second one, HRnet. In comparison, the lowest performer, the OpenPose, HRCSTrans achieves 17.9 points gain (89.2 compared to 71.3). In AR evaluation, RSN outperforms HRnet, securing the second position, while OpenPose remains the least effective with a score of 68.5. HRCSTrans secures the top position, boasting a 2.6 points gain over RSN (84.9 compared to 82.3). Across all the AP evaluation criteria, the HRCSTrans claims the first position, exhibiting a minimum 2.3 points gain (91.6 compared to 89.3). All the other backbones follow a similar distribution trend, except in the AP^{75} situation, where the RSN performs better than HRnet and reaches the second position. Compared to other OKS, all the keypoint deep learning networks are sensitive to the OKS = 0.5 situation and gain the highest scores. To provide a detailed perspective, Fig. 10 visualizes the specifics of AP^{50}. The AP value represents the ratio of the area under the Precision–Recall curve to the ideal area (set to be one). A more enormous AP value signifies an optimal performance of the deep learning network. It is evident that the HRCSTrans AP curve is closest to the top right corner of the recall/precision coordinates system. This positioning indicates that the red curve (HRCSTrans) has the largest area under the curve, emphasizing the robust keypoint detection performance of HRCSTrans in the bolt connection detection task, characterized by both high accuracy and stability.

Table 4
Ablation study.

Method	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
HR+CONV	86.1	87.9	83.1	82.8	83.3	80.9
HR+MHSA+FF	87.2	89.7	84.3	84.1	85.5	82.6
HR+MHSA+MSF	87.3	89.5	85.1	84.6	85.3	82.9
HR+DS-MHSA+FF	88.1	90.3	86.2	84.9	86.3	83.5
HR+DS-MHSA+MSF	**89.2**	**91.6**	**87.1**	**86.1**	**87.8**	**84.9**

In this work, the DS-MHSA is designed to replace MHSA in the vanilla transformer, while MSF is designed to replace the feedforward (FF) module. To prove the validity of DS-MHSA and MSF, an ablation study is depicted in Table 4. The hybridization of high-resolution architecture (HR) with convolution (CONV) or transformer is compared. The HR+MHSA or HR+DS-MHSA works better than HR+CONV, gaining at least 1.1 points on ap (87.2 compared to 86.1), illustrating that using the transformer to replace convolution in HR architecture will improve keypoint detection on bolt loosening evaluation tasks. The combination of HR+DS-MHSA+MSF reaches 89.2 AP and works best. The introduction of DS-MHSA gains 1.9 points (89.2 compared to 87.3) and MSF gets 1.1 points (89.2 compared to 88.1). The ablation study illustrates that the proposed DS-MHSA and MSF can effectively improve the precision of bolt images' keypoint detection.

Additionally, to assess the feasibility of implementing the HRCSTrans in industrial environments, the computational cost is evaluated and compared with other baselines. Three criteria are considered: the number of model parameters, floating point operations (FLOPs), and testing time per image. Table 5 presents the evaluation results for the baselines in terms of these criteria. Benefiting from the delicate convolution and branchless structure, RSN achieves the minimum number of parameters and FLOPs. The HRNet exhibits a 10% increase in parameters compared to RSN due to its utilization of the high-resolution multi-branch architecture. Our HRTrans, which replaces convolution with modified transformer blocks, experienced a 6% parameter increase compared to HRNet. The global feature aggregation of the transformer takes up more computing resources, leading to a 50% increase in FLOPs (24.7 compared to 16.0). Although the HRTrans has larger model parameters, its testing time ranks among the top three, being only 0.42 s slower than RSN and 0.28 s slower than HRNet. The proposed method can realize the prediction in only 0.55 s, which is in the same order of magnitude as the fastest RSN and demonstrates sufficient industrial application value.

To further demonstrate the network performance of HRCSTrans, the visualization of the location distribution heatmaps is shown in

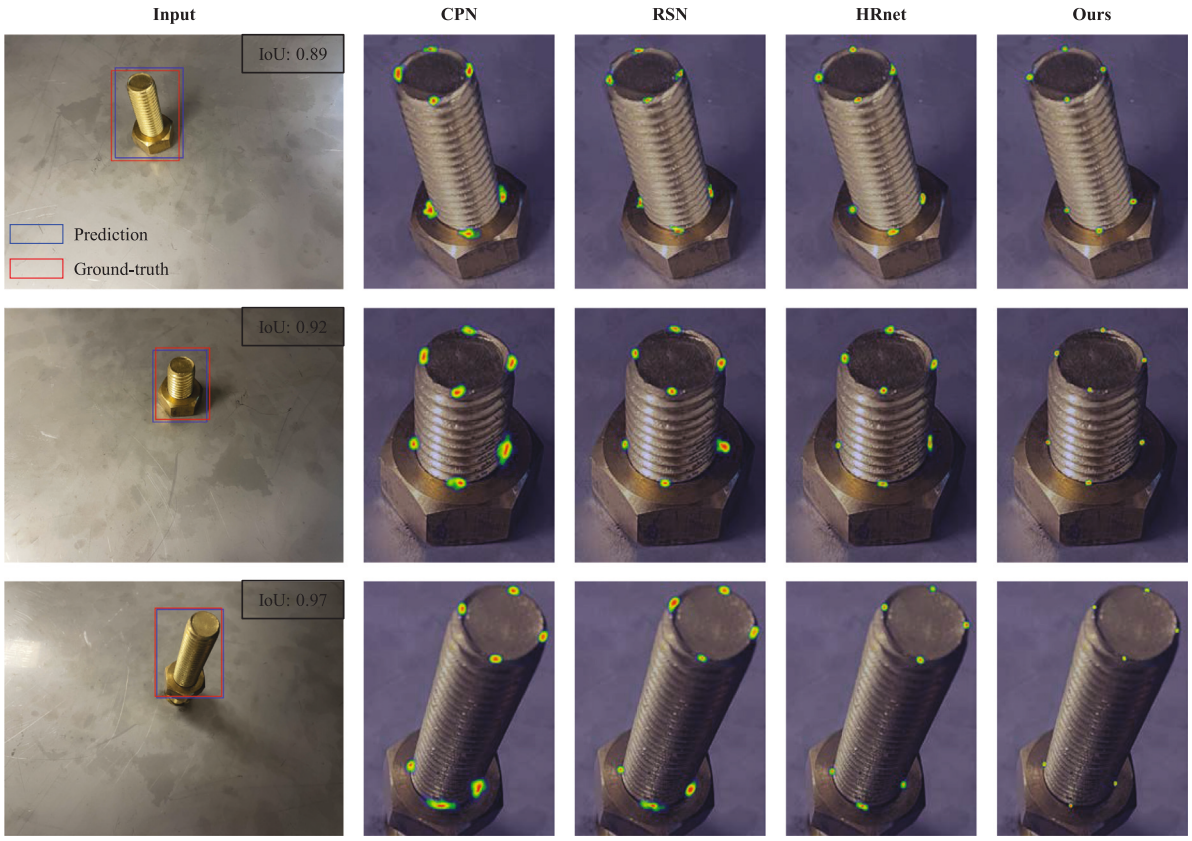


Fig. 11. Visualization of HRCSTrans heatmaps.

Table 5
Computational cost evaluation.

Model	Parameters (M)	FLOPs (G)	Testing time (s)
OpenPose	25.94	160.36	2.02
G-RMI	42.6	57.0	1.21
CPN	58.8	29.2	1.03
SimpleBase	68.6	35.6	0.87
RSN	25.7	6.4	0.13
HRnet	28.5	16.0	0.27
Ours	**30.1**	**24.7**	**0.55**

Fig. 11. The left column is the input image tested by the keypoint detection network. The blue line is the bounding box detected by RetinaNet used as ROI, while the red line is the labeled ground truth. The RetinaNet achieves 0.91 mean intersection over union (IoU) on the bolt connection dataset. The IoU of these inputs are indicated in Fig. 11. All the IoU are larger than 0.85, meaning the RetinaNet extracts the ROI with an acceptable performance for the keypoint detection. In addition to our method, we also visualize the top two performers among other methods, the HRnet and RSN. Additionally, the keypoint detection diagram, CPN, is also visualized. Compared to other methods, the heatmap area of HRCSTrans is obviously concentrated, meaning that the HRCSTrans has better network convergence and higher prediction accuracy. Keypoints F and G are challenged to detect due to shadow interference, and the phenomenon can be seen in the heatmap. The heatmap area of F and G in the CPN and RSN is larger than the area of other keypoints. The heatmap of keypoint F in HRnet is larger than the other six keypoints, indicating the difficulty of detecting keypoint F. Conversely, HRCSTrans exhibits consistent and smallest heatmap areas across different keypoints, illustrating robust anti-interference and stability in our method.

4.3. Monocular camera measurement result

Utilizing the keypoints detection network, the 2D image coordinates are obtained. The image coordinate (u, v) and the camera parameters obtained in Section 4.1 are transmitted to Eqs. (4)–(15), and the predicted exposed length is calculated. Comparing these prediction lengths to the ground truth measured by CMMS (Fig. 8), the method error is evaluated. Randomly select two samples from each additional ten-centimeter exposure length interval. The results are tabulated in Table 6, revealing a maximum error of 0.074 mm, with none exceeding 0.1 mm and a minimum error of only 0.008 mm. The heatmaps of these ten cases are shown in Fig. 12. The small heatmap areas detected by the HRCSTrans indicate excellent keypoint detection performance.

To further illustrate the precision of our method, we also compared the exposed length measurement results with other deep learning networks, as depicted in Fig. 13. Among all the seven methods, OpenPose demonstrates the poorest performance with 0.632 mm error and 0.151 mm standard deviation (STD), while the HRnet reaches the top one accuracy. The error distribution is linked to the AP value of keypoint detection. Higher AP points indicate lower errors and STD. HRCSTrans performs best during all the keypoint detection networks, exhibiting a 0.053 mm error and a 0.028 mm STD. This represents a 0.061 mm reduction in error compared to the second-best, HRnet (0.053 compared to 0.114), and a 0.030 lower STD (0.028 compared to 0.058). Empirical results from the experiment demonstrate that our proposed method exhibits superior precision, coupled with minimal error variability and enhanced stability. The average error of 0.053 mm adding the maximum float is lower than 0.1 mm, meaning that our approach satisfies the high precision assembly assurance requirements of the industrial sector, ensuring reliable and accurate performance. The vision-based method adapts to dynamic environmental changes and keeps high accuracy under long-term multiple excitations. The monocular measurement has been conducted under different vibration

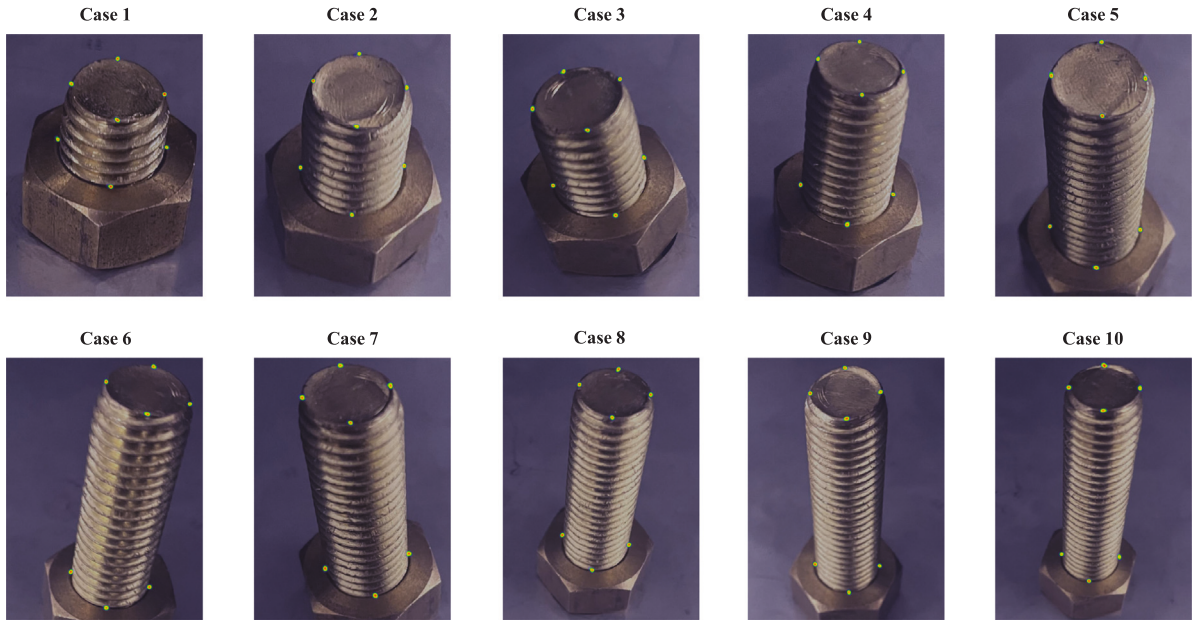


Fig. 12. Heatmaps of ten cases.

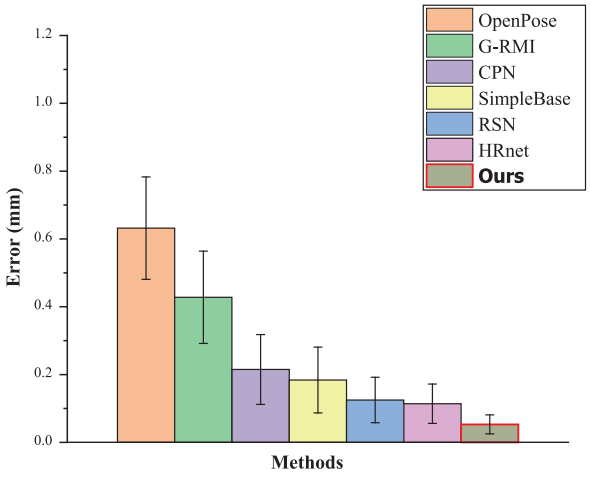


Fig. 13. Measurement results compare with other methods.

frequencies ranging from 500 to 2000 Hz. The hourly measured data across ten hours shows no significant differences with static measurement. Similarly, experimental testing at controlled temperature from 0 to 60 degrees Celsius consistently showed low measurement errors (<0.070 mm), indicating reliable performance under varying conditions. The measurement accuracy of 0.053 mm also enables the proposed vision system to observe slight changes in the connecting part, providing the possibility to measure changes in prestress through visual measurement. We conducted measurements for bolt connections with prestress loading under 10 Nm, 15 Nm, 20 Nm, and 25 Nm initial torques. The bolt length measured by the monocular vision also showed strong correlations with the prestress measurement with small mean absolute percentage errors. The monocular vision measurement model realizes spatial distance calculation in 0.14 s, while the HRCSTrans takes 0.55 s for an image prediction. The proposed bolt connection loosening detection system spends 0.69 s for an image evaluation, satisfying the real-time monitoring requirement.

Table 6
3D construction results and comparison with ground truth.

	1	2	3	4	5
3D construction	12.674 mm	18.997 mm	20.547 mm	26.035 mm	35.165 mm
Ground truth	12.659 mm	19.033 mm	20.584 mm	26.109 mm	35.142 mm
Error	0.015 mm	0.036 mm	0.037 mm	0.074mm	0.023 mm
	6	7	8	9	10
3D construction	38.226 mm	41.274 mm	45.741 mm	51.445 mm	52.289 mm
Ground truth	38.277 mm	41.234 mm	45.733 mm	51.376 mm	52.261 mm
Error	0.051 mm	0.040 mm	0.008 mm	0.069 mm	0.028 mm

4.4. System robustness test

In the preceding experiment, the evaluation of keypoints extraction and 3D construction results substantiated the effectiveness and high precision of our method. To further demonstrate the robustness of the proposed method, the image acquisition is implemented in diverse working situations, including different bolt materials, low brightness lighting conditions, multiple objects, unclear backgrounds, and different standard sizes, as shown in Fig. 14. For each situation, two examples are provided, each featuring two graphs. One graph presents the keypoint detection outcomes and the measurement error of the bolt's exposed length, as computed by the 3D construction model and CMM ground truth. The second graph displays the heatmap of the corresponding target, with a local magnification added to elucidate the visual details of keypoint detection results.

As for the different materials situation, the bolts made of stainless steel and alloy steel are tested by our method. The measurement error of the stainless steel bolt is 0.041 mm, similar to the experiment results on the brass bolt, and the keypoint of the heatmap is well converged. Due to the black color weakening the boundary between the bolt and nut, the keypoint detection performance on alloy steel bolts is worse than the stainless and brass. The heatmap of the area of keypoints is obviously more extensive, and the error increases to 0.087 mm. Similar to the alloy steel bolt, the keypoint extraction results deteriorate in the low brightness lighting condition situation, leading to increased measurement errors in two examples (0.075 mm and 0.082 mm). Nevertheless, the measurement errors, even with weakened boundaries, remain below 0.1 mm. The errors are significantly smaller than the actual exposed bolt length, showcasing at least a 200-fold

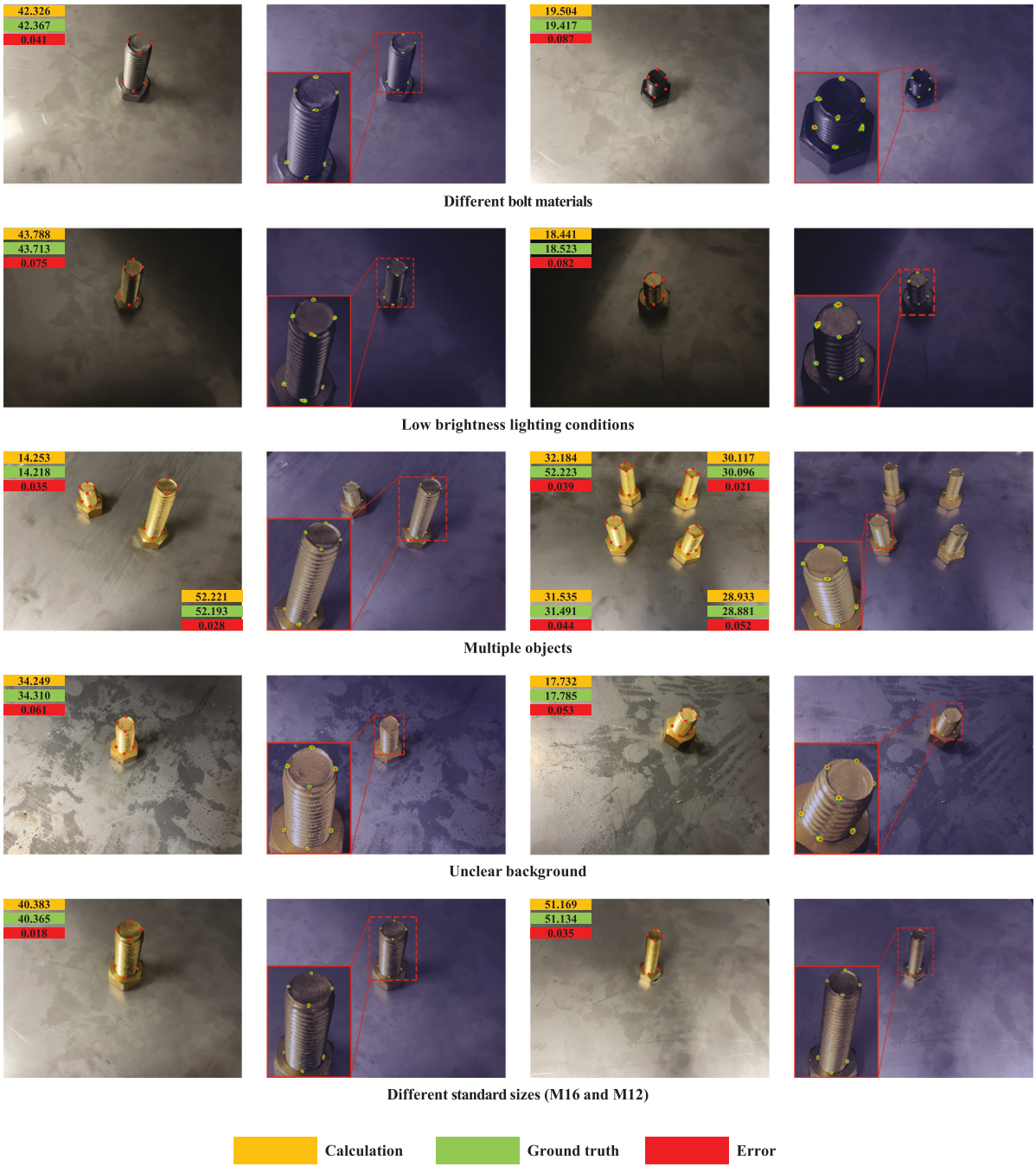


Fig. 14. Robustness test.

difference on the length scale. In the multiple objects situation, two kinds of bolt arrays (two and four objects) are tested and yield an average error of 0.0365 mm, with a maximum error of 0.052 mm. Compared to the average error illustrated in Fig. 13 (0.053 mm), HRCSTrans performance on the multiple objects situation is very stable and accurate. To simulate a real industrial working environment, paint marks are added to the background, increasing the difficulty of image processing. Despite the complex background, HRCSTrans is still stable, with 0.061 mm and 0.053 mm errors, respectively. There is no apparent fluctuation in the heatmap results. Finally, bolts of different standard sizes (M16 and M12) are tested on the HRCSTrans. The measurement errors are only 0.018 mm and 0.035 mm, illustrating our method is generalized and it can be adapted to different bolt sizes. In summary, the measurement precision remains consistent in the multiple objects, unclear background and different bolt standard sizes situations. As for low brightness lighting conditions and alloy steel materials, where image boundary blurring occurs, there is a slight increase in measurement errors (no more than 0.05 mm). However, this error increase is negligible relative to the bolt exposed length (0.05 mm compared to 30 mm).

To further illustrate the robustness of the proposed method, an outside bolt connections dataset is constructed to test the HRCSTrans. The outside dataset consists of 150 images, with 75 images allocated for training and the remaining 75 images for testing. The parameters of HRCSTrans are transferred from the training results in Section 4.2 and then trained using the outside dataset for 500 epochs. The outside dataset encompasses various situations, including double nuts, the corrosion surface, paint coating, and rainy weather conditions. The measurement results and corresponding heatmaps for these cases are presented in Fig. 15. The ground truth is measured by the vernier scale,

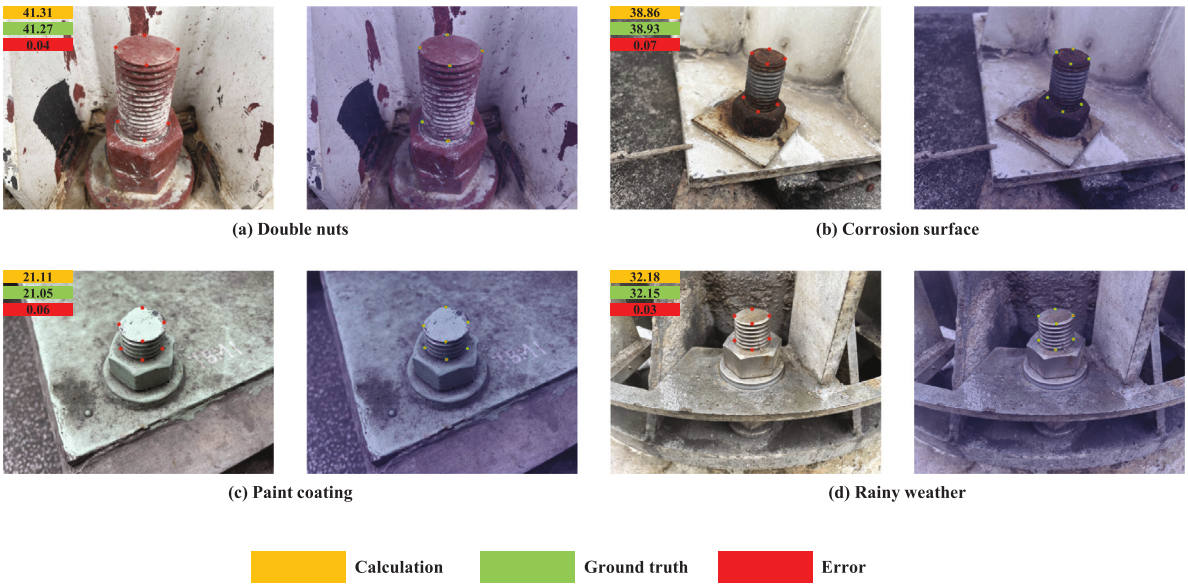


Fig. 15. Outside bolt connections test.

and all the prediction errors are lower than 0.1 mm, which is consistent with the indoor test accuracy. The convergence performance of the heatmaps demonstrates the robustness of HRCSTrans in handling outside bolt connections. The robustness test affirms that our method can adapt to diverse working situations and target objects, demonstrating practical value in natural industrial environments.

4.5. Discussion

In the experiment section, an image acquisition platform featuring various bolts is established. Following the collection of image information, the bolt being measured is transported to a high-precision CMM to ascertain the ground truth of the exposed length. All the images, with annotated keypoints and matching ground truth, are used to construct a deep learning data set. Our HRCSTrans and other keypoint detection deep learning baselines are tested on this dataset. Then, the monocular vision measurement model calculates the exposed bolt length based on the keypoint extraction results with the camera calibration parameters. The experiment results reveal that our HRCSTrans realized 91.6 AP^{50} prediction points and AR 84.9 points, positioning it as the SOTA method in the bolt keypoint detection task. The measurement error of the monocular camera model is only 0.053 mm with 0.028 mm STD. Furthermore, our measurement method is tested in five different industrial situations. The measurement performance consistently maintains stability, with errors not exceeding 0.1 mm. This illustrates the high robustness of our measurement method, demonstrating its adaptability to effectively to the actual manufacturing environment. The use of transformer mechanisms ensures the capacity of the proposed model for large training data sets, increasing the potential for large-scale industrial applications.

In future work, the distribution of data will be expanded to account for more real-world scenarios, which aims to improve the reliability of the method. Transfer learning and domain adaptation will be used to reduce training costs and training data requirements. Additionally, the implementation of monocular vision limits the system's depth perception capabilities. The proposed system will be combined with the depth images to broaden the application scope of bolt connection assessment. As for the deep learning model, this work exclusively evaluates a singular model size. Multiple model sizes employing the proposed backbone need to be developed to address the diverse data sizes inherent in various industrial scenarios. Moreover, the proposed vision system is dedicated to achieving high-precision measurement and detecting the loosening state based on exposed length. There is potential for further exploration of the relationship between bolt connection prestress and exposed length. Specifically, the bolt connection prestress can be recorded as a predictive label upon acquiring the bolt image. This integration facilitates prestress prediction during the exposed length calculation process.

5. Conclusion

In this paper, a keypoint detection deep learning network and monocular vision measurement model hybrid system is proposed to realize noncontact bolt measurement and connection loosening detection. To obtain high precision keypoint extraction results, a new backbone, HRCSTrans, is proposed, incorporating the transformer mechanism into vision-based bolt connection loosening detection. To save the working space on the image acquisition system, a monocular vision measurement model is proposed to construct the 3D model of bolts and evaluate the loosening state. Compared to contact monitoring methods, the vision-based method can keep the integrity of the bolt connection and save the time cost of data collection time. Additionally, different from the force-sensor based measurement that requires tight connections, the vision-based sensor can detect various loosening situations, providing a wide range of measurements. The experiment results show that HRCSTrans realizes the top one AP value and AR score compared with other baselines. The ablation study verified the effectiveness of the transformer and high-resolution hybridized architecture. The overall measurement system reaches 0.053 mm precision and can adapt to different dataset domains, indicating the deployment feasibility in real industrial situations. The primary contributions of this study include:

(1) This is the first attempt to introduce the transformer mechanism in the bolt keypoint detection for connection loosening detection. The transformer deep learning model provides an approximate global view and extracts high precision keypoints, guaranteeing the three-dimension construction of bolt surfaces.

(2) A novel keypoint detection backbone, HRCSTrans, is proposed for effective keypoint feature learning. To further improve the baseline performance, the transformer block is inserted into the multiple resolutions architecture for aggregating affluent inter-level features. The lightweight patch embedding and cross-scale transformer block are designed to prevent the computational explosion caused by the mix of multiple resolutions architecture and transformer block. To get better intra-level information fusion, the dual-scale multi-head self-attention

and multi-scale feedforward block are designed for cross patch feature aggregation. The experiment depicts HRCSTrans realizing the top AP and AR value in the bolt keypoint detection tasks. The convergence effect of the heat map also verifies the stability and reliability of the novel backbone.

(3) A new monocular 3D construction model based on 2D keypoints coordinates is proposed. In contrast to the multi-vision system, the proposed monocular model saves the cost of constructing multi-camera coordinate system associations and minimizes the workspace in image acquisition platform configuration. Moreover, this method accurately calculates the 3D scale information of bolts without the need to attach additional marks.

(4) The proposed measurement system achieves a fully automated quantitative detection of the bolt connection loosening state. It can reach 0.1 mm accuracy measurement, meeting industrial needs and exceeding existing detection methods by an order of magnitude. The method demonstrates robust performance across various industrial situations, including different bolt materials, low brightness lighting conditions, multiple objects, unclear backgrounds, and different standard sizes.

The monocular vision and deep learning hybrid bolt measurement system address the connection loosening detection problem. With trained networks and the monocular vision system, the exposed bolt length can be calculated without attaching sensors or marks, providing quantitative detection of connection statements. Experiment results illustrate that HRCSTrans achieves SOTA performance in the bolt keypoint detection tasks and affirm the precision of the monocular vision measurement model. The system's robustness is validated across diverse working situations, confirming its deplorability in industrial environments. In the future, the image acquisition scenarios will be expanded to illustrate the method's generalization capabilities.

CRediT authorship contribution statement

Wu Tianyi: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shang Ke:** Visualization, Conceptualization. **Dai Wei:** Validation, Conceptualization. **Wang Min:** Validation, Conceptualization. **Liu Rui:** Validation, Conceptualization. **Zhou Junxian:** Validation, Conceptualization. **Liu Jun:** Supervision, Project administration, Investigation, Funding acquisition.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgments

This work was supported by the Research Grant Council (RGC) of Hong Kong under Grant 11217922, 11212321 and Grant ECS-21212720, and the Science and Technology Innovation Committee of Shenzhen under Grant Type-C SGDX20210823104001011.

References

Ali, R., Cha, Y.J., 2022. Attention-based generative adversarial network with internal damage segmentation using thermography. Autom. Constr. 141, 104412.
Ali, R., Kang, D., Suh, G., Cha, Y.J., 2021. Real-time multiple damage mapping using autonomous UAV and deep faster region-based neural networks for GPS-denied structures. Autom. Constr. 130, 103831.
Bertasius, G., Wang, H., Torresani, L., 2021. Is space-time attention all you need for video understanding? In: ICML, vol. 4.
Cai, Y., Wang, Z., Luo, Z., Yin, B., Du, A., Wang, H., Zhang, X., Zhou, X., Zhou, E., Sun, J., 2020. Learning delicate local representations for multi-person pose estimation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. Springer, pp. 455–472.
Cha, Y.J., Choi, W., Büyüköztürk, O., 2017. Deep learning-based crack damage detection using convolutional neural networks. Comput.-Aided Civ. Infrastruct. Eng. 32, 361–378.
Cha, Y.J., Choi, W., Suh, G., Mahmoudkhani, S., Büyüköztürk, O., 2018. Autonomous structural visual inspection using region-based deep learning for detecting multiple damage types. Comput.-Aided Civ. Infrastruct. Eng. 33, 731–747.
Cha, Y.J., You, K., Choi, W., 2016. Vision-based detection of loosened bolts using the hough transform and support vector machines. Autom. Constr. 71, 181–188.
Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., 2018b. Cascaded pyramid network for multi-person pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7103–7112.
Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018a. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European Conference on Computer Vision. ECCV, pp. 801–818.
Choi, W., Cha, Y.J., 2019. Sddnet: Real-time crack segmentation. IEEE Trans. Ind. Electron. 67, 8016–8025.
Dong, X., Bao, J., Chen, D., Zhang, W., Yu, N., Yuan, L., Chen, D., Guo, B., 2022. Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12124–12134.
Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2010. An image is worth 16x16 words: Transformers for image recognition at scale. arxiv 2020. arXiv preprint arXiv:2010.11929.
Duan, C., Zhang, H., Li, Z., Tian, Y., Chai, Q., Yang, J., Yuan, L., Zhang, J., Xie, K., Lv, Z., 2019. FBG smart bolts and their application in power grids. IEEE Trans. Instrum. Meas. 69, 2515–2521.
Feng, H., Jiang, Z., Xie, F., Yang, P., Shi, J., Chen, L., 2013. Automatic fastener classification and defect detection in vision-based railway inspection systems. IEEE Trans. Instrum. Measur. 63, 877–888.
Gong, H., Deng, X., Liu, J., Huang, J., 2022. Quantitative loosening detection of threaded fasteners using vision-based deep learning and geometric imaging theory. Autom. Constr. 133, 104009.
Gu, J., Kwon, H., Wang, D., Ye, W., Li, M., Chen, Y.H., Lai, L., Chandra, V., Pan, D.Z., 2022. Multi-scale high-resolution vision transformer for semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12094–12103.
Hei, C., Luo, M., Gong, P., Song, G., 2020. Quantitative evaluation of bolt connection using a single piezoceramic transducer and ultrasonic coda wave energy with the consideration of the piezoceramic aging effect. Smart Mater. Struct. 29, 027001.
Hong, D., Yokoya, N., Chanussot, J., Zhu, X.X., 2018. An augmented linear mixing model to address spectral variability for hyperspectral unmixing. IEEE Trans. Image Process. 28, 1923–1938.
Hosseinpour, M., Daei, M., Zeynalian, M., Ataei, A., 2023. Neural networks-based formulation for predicting ultimate strength of bolted shear connectors in composite cold-formed steel beams. Eng. Appl. Artif. Intell. 118, 105614.
Jamil, S., Roy, A.M., 2023. An efficient and robust phonocardiography (PCG)-based valvular heart diseases (VHD) detection framework using vision transformer (VIT). Comput. Biol. Med. 158, 106734.
Jiang, B., Chen, S., Wang, B., Luo, B., 2022. Mglnn: Semi-supervised learning via multiple graph cooperative learning neural networks. Neural Netw. 153, 204–214.
Kang, D.H., Cha, Y.J., 2022. Efficient attention-based deep encoder and decoder for automatic crack segmentation. Struct. Health Monit. 21, 2190–2205.
Lewis, J., Cha, Y.J., Kim, J., 2023. Dual encoder–decoder-based deep polyp segmentation network for colonoscopy images. Sci. Rep. 13, 1183.
Lin, T.Y., Dollár, R., He, K., Hariharan, B., Belongie, S., 2017a. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017b. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988.
Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, C.L., 2014. Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, pp. 740–755.
Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.

Ma, D., Fang, H., Wang, N., Lu, H., Matthews, J., Zhang, C., 2023. Transformer-optimized generation, detection, and tracking network for images with drainage pipeline defects. Comput.-Aided Civ. Infrastruct. Eng..

Mazzeo, P.L., Nitti, M., Stella, E., Distante, A., 2004. Visual recognition of fastening bolts for railroad maintenance. Pattern Recognit. Lett. 25, 669–677.

Miao, R., Shen, R., Zhang, S., Xue, S., 2020. A review of bolt tightening force measurement and loosening detection. Sensors 20, 3165.

Mushtaq, F., Ramesh, K., Deshmukh, S., Ray, T., Parimi, C., Tandon, P., Jha, P.K., 2023. Nuts&bolts: Yolo-v5 and image processing based component identification system. Eng. Appl. Artif. Intell. 118, 105665.

Ramana, L., Choi, W., Cha, Y.J., 2019. Fully automated vision-based loosened bolt detection using the Viola–Jones algorithm. Struct. Health Monit. 18, 422–434.

Ren, L., Feng, T., Ho, M., Jiang, T., Song, G., 2018. A smart "shear sensing" bolt based on FBG sensors. Measurement 122, 240–246.

Rosso, M.M., Aloisio, A., Randazzo, V., Tanzi, L., Cirrincione, G., Marano, G.C., 2023. Comparative deep learning studies for indirect tunnel monitoring with and without fourier pre-processing. Integr. Comput.-Aided Eng. 1–20.

Wang, F., 2023. Multi-bolt looseness detection using a new acoustic emission strategy. Struct. Health Monit. 22, 1543–1553.

Wang, F., Chen, Z., Song, G., 2020a. Monitoring of multi-bolt connection looseness using entropy-based active sensing and genetic algorithm-based least square support vector machine. Mech. Syst. Signal Process. 136, 106507.

Wang, Z., Liu, M., Zhu, Z., Qu, Y., Wei, Q., Zhou, Z., Tan, Y., Yu, Z., Yang, F., 2019b. Clamp looseness detection using modal strain estimated from FBG based operational modal analysis. Measurement 137, 82–97.

Wang, F., Song, G., 2019. Bolt early looseness monitoring using modified vibro-acoustic modulation by time-reversal. Mech. Syst. Signal Process. 130, 349–360.

Wang, J., Sun, K., Cheng, T., Jiang, B., Deng, C., Zhao, Y., Liu, D., Mu, Y., Tan, M., Wang, X., et al., 2020b. Deep high-resolution representation learning for visual recognition. IEEE Trans. Pattern Anal. Mach. Intell. 43, 3349–3364.

Wang, T., Tan, B., Lu, G., Liu, B., Yang, D., 2020c. Bolt pretightening force measurement based on strain distribution of bolt head surface. J. Aerosp. Eng. 33, 04020034.

Wang, C., Wang, N., Ho, S.C., Chen, X., Song, G., 2019a. Design of a new vision-based method for the bolts looseness detection in flange connections. IEEE Trans. Ind. Electron. 67, 1366–1375.

Wei, D., Wei, X., Tang, Q., Jia, L., Yin, X., Ji, Y., 2023. Rtlseg: A novel multi-component inspection network for railway track line based on instance segmentation. Eng. Appl. Artif. Intell. 119, 105822.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. Adv. Neural Inf. Process. Syst. 34, 12077–12090.

Yang, F., Zhang, L., Yu, S., Prokhorov, D., Mei, X., Ling, H., 2019. Feature pyramid and hierarchical boosting network for pavement crack detection. IEEE Trans. Intell. Transp. Syst. 21, 1525–1535.

Zhang, Z., 1999. Flexible camera calibration by viewing a plane from unknown orientations. In: Proceedings of the Seventh Ieee International Conference on Computer Vision. Ieee, pp. 666–673.

Zhang, P., Dai, X., Yang, J., Xiao, B., Yuan, L., Zhang, L., Gao, J., 2021. Multi-scale vision longformer: A new vision transformer for high-resolution image encoding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2998–3008.

Zhang, X., Yan, Q., Yang, J., Zhao, J., Shen, Y., 2019. An assembly tightness detection method for bolt-jointed rotor with wavelet energy entropy. Measurement 136, 212–224.

Zhao, K., Wang, Y., Zuo, Y., Zhang, C., 2022. Palletizing robot positioning bolt detection based on improved Yolo-V3. J. Intell. Robot. Syst. 104 (41).

Zhao, X., Zhang, Y., Wang, N., 2019. Bolt loosening angle detection technology using deep learning. Struct. Control Health Monit. 26, e2292.

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6881–6890.