Measurement 240 (2025) 115582

Contents lists available at ScienceDirect

Measurement

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

An algorithm for large-span flexible bridge pose estimation and multi-keypoint vibration displacement measurement

Ruiyang Sun[1], Sen Wang *, Mao Li[1], Yang Zhu[1]

Faculty of Mechanical and Electrical Engineering, Kunming University of Science and Technology, Kunming 650500, China
Key Laboratory of Advanced Equipment Intelligent Manufacturing Technology of Yunnan Province, Kunming 650500, China

ARTICLE INFO

Keywords:
Visual vibration measurement
Flexible structures
Bridge keypoint tracking
Deep convolutional neural network

ABSTRACT

Traditional visual vibration measurement methods face limitations in dynamic outdoor environments, making accurate multi-point vibration measurements challenging for the targeted objects. In this paper, a visual vibration measurement algorithm for large-span bridges is proposed by anthropomorphizing the bridge target and introducing key point detection at the same time. Firstly, the algorithm utilizes a convolutional neural network to extract multi-scale feature information from the image sequence of the bridge model. Secondly, to enhance target detection accuracy, a coordinate attention (CA) mechanism is incorporated, and the shape intersection over union (SIoU) loss function replaces the original loss function. Finally, the algorithm integrates the density-based spatial clustering of applications with noise (DBSCAN) and tracking algorithms to achieve precise localization of bridge targets. Vibration displacement data were extracted from a large-span suspension bridge structure in an outdoor environment and analyzed quantitatively and qualitatively. The vibration displacement curves obtained in this study show the highest level of agreement with the standard displacement signals, with a mean absolute percentage error of 0.5170% for the cable-stayed bridge model, 1.3822% for the Humen Bridge, and 3.2263% and 1.6982% for the Longjiang Bridge.

1. Introduction

In recent decades, there has been a significant global surge in the construction of large-span suspension bridge. However, with the increase in the main spans and tower heights of these bridges, these flexible structures are more susceptible to experiencing overall oscillations when subjected to dynamic loads such as strong winds, earthquakes, and vehicular movements [1]. This long-term, high-amplitude, low-frequency vibration mode is bound to induce structural fatigue, ultimately leading to the failure of bridge structures. For instance, on November 7, 1941, the Tacoma Narrows Bridge in Washington State underwent oscillations when exposed to wind speeds of 18 m/s, resulting in structural failure caused by vibration-induced collapse. Similarly, on May 5, 2020, China's Humen Bridge in Guangzhou encountered vortex-induced vibrations due to specific wind flow patterns, leading to the closure of the bridge's lanes. Consequently, the real-time monitoring of vibrations in flexible structures like bridges and the assessment of structural damage during their service life based on structural responses hold substantial importance [2].

In the realm of structural vibration measurement, there are primarily two approaches: contact-based and non-contact-based methods. Traditional contact-based vibration measurement techniques have demonstrated their efficacy within specific scenarios. However, their precision in capturing vibration modes and frequencies can be compromised due to variations and mutual interference in attributes like object stiffness, structural characteristics, and quality [3–5]. For instance, linear variable differential transformers (LVDT) require a highly controlled installation environment, necessitating a fixed position as a reference point and physical connections to data acquisition and power supply equipment. This poses a considerable challenge, particularly in the context of complex and dynamic bridge environments [6]. As a result, researchers have begun to focus their attention on the development of non-contact vibration measurements such as GPS vibrometry, millimeter-wave radar vibrometry, and laser vibrometry. Ma et al. [7] proposes a bridge displacement estimation technique that combines an accelerometer, a strain gauge, and a millimeter-wave radar considering intermittent radar target occlusion common in long-term displacement monitoring. Geng et al. [8] established a new system for contactless continuous measurement of blood pressure based on a single millimeter-wave radar for simultaneous acquisition of the central arterial pulse transit time in the carotid and thoracic regions, and validated the potential of this system for continuous blood

* Correspondence to: 727 South Jingming Road, Chenggong District, 650500 Kunming, China.
E-mail addresses: sry3020037@163.com (R. Sun), wangsen0401@126.com (S. Wang), limao_w@163.com (M. Li), zhuyang1023@foxmail.com (Y. Zhu).
[1] 727 South Jingming Road, Chenggong District, 650500, Kunming, China.

https://doi.org/10.1016/j.measurement.2024.115582
Received 20 April 2024; Received in revised form 1 August 2024; Accepted 21 August 2024
Available online 24 August 2024
0263-2241/© 2024 Elsevier Ltd. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

pressure measurement. Zhang et al. [9] proposed a new method based on microwave interferometry and rotational influence line theory to accurately identify the stiffness distribution of continuous girder bridges, and verified the validity of this method to identify the stiffness distribution of continuous girder bridges through laboratory experiments. GPS sensors offer a sampling frequency of up to 20 Hz. However, they are susceptible to external factors such as electromagnetic interference and atmospheric conditions, especially in the vertical direction where their measurement accuracy is relatively poor [10–12]. The wavelengths used in millimeter-wave lidar are at the millimeter level, which leads to excessive roughness requirements on the measurement target, greater influence by complex environments, and shorter measurement distances. Conversely, Doppler laser vibrometers are vulnerable to factors such as temperature, humidity, dust, and surface reflections from the measured objects, leading to reduced measurement precision [13].

With the advancement of image processing algorithms and the enhanced performance of image acquisition devices, the field of vision-based non-contact measurement techniques has witnessed experienced notable progress. This technology has garnered effective validation in the realm of structural deformation monitoring owing to its capability to real-time monitor the vibration status of multiple targets on bridges over substantial distances [14–19]. Habeenzu et al. used an unmanned aerial system (UAS) for data acquisition of bridges and visual measurements with sub-millimeter accuracy. Shang et al. [20] introduced a multi-point vibration measurement approach based on optical flow, which they validated on cantilever beams and steel bridges. Zhang et al. [21] proposed a non-contact vibration displacement measurement method for rotating axes based on Gaussian fitting and edge optimization, and conducted practical experiments using a turbine sensor to verify the accuracy of the method. Song et al. [22] propose a robust respiratory rate (RR) measurement method using a two-level fusion of video and FMCW (frequency modulated continuous wave) radar information. This RR measurement method works significantly better than the single mode fusion method. Liu et al. [23] proposed a phase-based optical flow method, enhancing vibration measurement accuracy through derivative operations. Yang et al. [24] proposed an improved method based on QR code marker tracking and phase mapping method on the basis of optical flow-based target tracking and detection algorithm, and verified the proposed improved scheme by theoretical calculation. In contrast to traditional vibration measurement methods, vision-based vibration measurement demonstrates commendable precision, a fact substantiated by these investigations and their corresponding findings [25–30].

Currently, the majority of vision-based vibration measurement methods rely on traditional digital image processing algorithms, which are often susceptible to environmental conditions and variations in target shapes. For instance, Feng et al. [31] employed an enhanced OCM template matching technique to track targets on bridge surfaces. The accuracy of template matching techniques is strongly contingent on image quality and is considerably compromised under conditions of changing illumination, target occlusion, and varying weather conditions. These factors impact the precision of target feature matching. Furthermore, during vibration processes, matched targets may experience rotation, deformation, and other alterations, further affecting the accuracy of visual vibration measurements [32,33]. Techniques like interpolation, parabolic fitting, and Gaussian fitting can improve the precision of template matching and displacement detection to sub-pixel levels [34]. However, they still remain susceptible to the influence of complex environmental factors, making accurate target detection challenging. Yoon et al. [15] proposed a data processing method based on optical flow estimation for target feature tracking, enhancing detection accuracy through the introduction of outlier detection and sub-pixel detection methods. Zhao et al. [35] combined Support Correlation Filters (SCF) with Kanade–Lucas–Tomasi (KLT), thereby enhancing the accuracy and robustness of vibration displacement measurement for cable-stayed bridges. Operating on the premise of brightness consistency, Guo et al. [36] introduced the Kanade–Lucas algorithm into a high-speed acquisition system and verified the system's accuracy and effectiveness in dynamic remote displacement measurement through experiments conducted on a moving platform on an elevated bridge and an acoustic barrier. Nonetheless, the Kanade–Lucas algorithm may experience reduced accuracy in real-world detection scenarios due to factors such as changes in illumination and image noise. Given the susceptibility of template matching algorithms and the Kanade–Lucas algorithm to environmental factors, it is crucial to choose for more reliable detection algorithms.

With the advancement of computer hardware and the enhancement of GPU computing power, the application of deep learning algorithms in fields such as image classification, object detection, and segmentation has gradually transitioned from performance comparisons on public datasets to structural engineering applications, including damage detection and vibration measurement [37–45]. Deep learning methods, in comparison to traditional techniques for visual vibration measurement, demonstrate robustness against background noise and detection environments. They effectively extract depth information from targets, resulting in improved precision and generalization in detection and tracking. This data-driven approach excels in feature extraction, reducing its reliance on acquisition conditions, brightness, and contrast. Guo et al. [38] introduced a visual structural measurement method that combines deep learning algorithms with optical flow, validated on the Wuyuan Bay Bridge in Xiamen, China. Both of these algorithms employ deep learning for assistance and subsequently extract vibration curves after detecting the target's position. However, these methods still suffer from the limitations of traditional algorithms, such as susceptibility to environmental factors and lower detection accuracy. Therefore, utilizing deep convolutional neural networks to extract target vibration displacement presents a promising strategy. For example, Lin et al. [39] proposed a high-precision displacement measurement algorithm based on deep convolutional neural networks, validated in a laboratory setting and under real outdoor bridge conditions. Currently, these deep learning algorithms primarily concentrate on detecting the vibration of individual targets, leaving substantial opportunities for advancement in terms of the diversity, accuracy, and stability of target position detection. This becomes particularly crucial when bridges experience flexible body vibrations, as the targets may undergo deviations and changes in orientation angles. This change affected the alignment of the target bounding boxes and the precision of target localization, thus emphasizing the importance of keypoint detection for detecting targets.

Inspired by visual detection techniques employed in the identification of joints in animals and the human body, we embarked on an endeavor illustrated in Fig. 1. Our aim is to conceptualize the entire structure of the flexible bridge in a manner reminiscent of an organism or the human physique. To achieve this, we transformed several identical targets (referred to as keypoints) located on the bridge into joint-like entities, evoking parallels with the articulations found in the human body or even in insects, such as flies. Through the precise capture of the positional variations of these multiple keypoints, we are able to quantify the deformation and displacement experienced by large-span bridges and other flexible structural entities.

Building on this anthropomorphic measurement approach, we convert the challenge of visual vibration measurement into the precise estimation of keypoint poses on large-span bridges. As a result, we introduce a deep learning-based algorithm designed for bridge pose estimation and multi-keypoint vibration displacement measurement. This algorithm effectively calculates the positional deviations of multiple keypoints by estimating the bridge's pose. Our research focuses on large-span bridges, and we initiate the process by training and validating the neural network model using a laboratory-scale cable-stayed bridge model. Subsequently, we evaluate the algorithm's performance

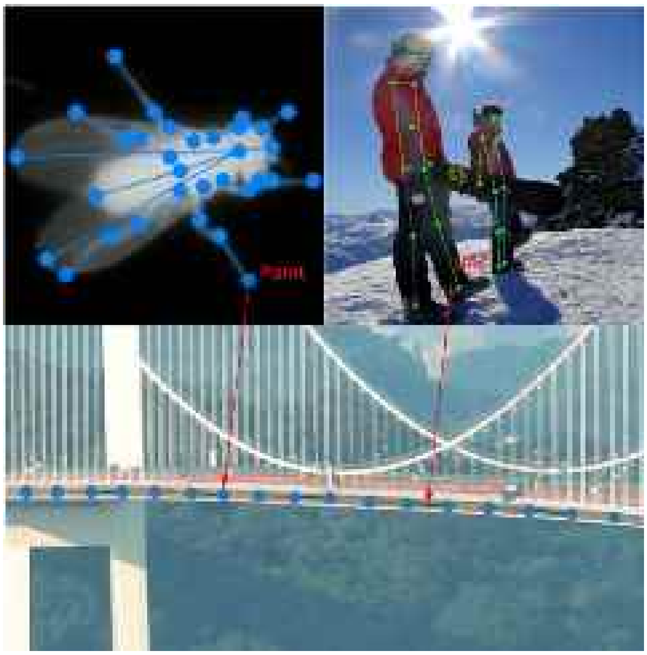

Fig. 1. Schematic representation of vibration measurement based on keypoint.

in a real-world scenario on a large-span suspension bridge. The implementation of our approach is depicted in Fig. 2, and the key innovations can be summarized as follows:

1. Given that bridges are subject to vibrational excitations, resulting in the occurrence of target vibration displacement or rotation, this algorithm represents a pioneering approach by integrating the pose information of keypoints related to detection targets. It departs from the conventional reliance solely on the target's central location for feature extraction. In contrast to algorithms that merely extract the central location of the target bounding box, this method exhibits a marked enhancement in detection accuracy.

2. Expanding on the YOLOv7-w6 framework, we incorporate the CA attention mechanism [46] to achieve precise target identification and localization. Furthermore, replacing the original loss function with the SIoU loss function [47] offers the dual benefit of reducing model training time while concurrently improving training efficacy.

3. In the process of model inference, we choose for the DBSCAN clustering algorithm over the Non-Maximum Suppression (NMS) algorithm. Clustering is applied to predictions characterized by high confidence scores, with the resultant positions of clustered class centers serving as the prediction outputs. This method effectively eliminates prediction noise, consequently elevating the accuracy of inference.

4. In order to augment the consistency of target position information across frames in video streams during inference and to mitigate the precision errors introduced by video noise in the process of extracting vibration targets, we employ the ByteTrack target tracking algorithm [48] to monitor the movement of keypoints associated with the targets. This application significantly improves the effectiveness of measuring vibration displacement.

This paper is structured as follows: Section 1 serves as an introduction, where we present the current state of visual vibration measurement and outline the relevant work undertaken in this study. Section 2 delves into a comprehensive explanation of the algorithms developed throughout this research. In Section 3, within the experimental section, we compare our algorithms' performance with mainstream deep learning algorithms, as well as displacement data acquired from accelerometer sensors. This comparative analysis substantiates the superiority of our approach. Section 4 extends our validation efforts. We apply our algorithms to the Guangdong Humen Bridge, subjected to vortex-induced vibrations, and the Yunnan Longjiang Bridge during its service life. In doing so, we conduct a thorough comparison of their performance against other algorithms. Lastly, Section 5 encapsulates our work, offering a summary of our findings and a forward-looking perspective on the contributions made in this paper.

Table 1
Composition of the backbone network model structure.

Layer	Name	From	Input size	Output size	K	S
0	ReOrg	–	640 × 640 × 3	320 × 320 × 12	–	–
1	CBS	−1	320 × 320 × 12	320 × 320 × 64	3	1
2	CBS	−1	320 × 320 × 64	160 × 160 × 128	3	2
3	ELAN	−1	160 × 160 × 128	160 × 160 × 128	–	–
4	CBS	−1	160 × 160 × 128	80 × 80 × 256	3	2
5	ELAN	−1	80 × 80 × 256	80 × 80 × 256	–	–
6	CBS	−1	80 × 80 × 256	40 × 40 × 512	3	2
7	ELAN	−1	40 × 40 × 512	40 × 40 × 512	–	–
8	CBS	−1	40 × 40 × 512	20 × 20 × 768	3	2
9	ELAN	−1	20 × 20 × 768	20 × 20 × 768	–	–
10	CBS	−1	20 × 20 × 768	10 × 10 × 1024	3	2
11	ELAN	−1	10 × 10 × 1024	10 × 10 × 1024	–	–

2. Algorithm structure

2.1. Framework for bridge keypoint pose estimation and detection network

The network framework for bridge keypoint pose estimation constructed in this paper consists of a backbone network and a head. The backbone network is responsible for extracting target features at different depths from input images. It achieves feature fusion and path enhancement for vibration target detection when these features are input into the head module. This process results in obtaining rich semantic information and accurate positional information. To further enhance the network's sensitivity to target positional information, we introduce the CA attention mechanism. Additionally, we improve the ByteTrack tracking algorithm to better suit keypoint tracking in flexible structures like bridges. The network model framework designed for multi-target keypoint pose estimation and detection on bridges is shown in Fig. 3.

The CSPDarknet53 neural network [49] serves as the fundamental network for extracting features in the context of bridge target analysis. The network primarily comprises ReOrg layers, ELAN layers, and convolutional layers with different kernel sizes. Significantly, the internal residual blocks in ELAN employ skip connections to address the common problem of gradient vanishing associated with deepening network architectures. Initially, the backbone network processes input images by segmenting and stacking them using the ReOrg layer, followed by feature extraction facilitated by the ELAN module. The resulting four feature maps are then directed into the head layer. The backbone structure of this study was as shown in Table 1.

2.2. Extraction and fusion of target features in bridge structures

In traditional CNN and FPN networks, the transmission of shallow-level information to deeper layers often follows a lengthy route, leading to the loss of intricate details in bridge target objects. To address this limitation, we incorporated a PANet [50] feature fusion structure, as shown in Fig. 4. That introduces a bottom-up enhancement path during the process of target feature extraction. This innovative architecture not only preserves fine-grained information from shallow-level networks but also effectively captures semantics from deeper networks, thereby markedly enhancing the precision of bridge target localization. To elaborate further, feature map P4 is derived from feature map Q4 through the utilization of the Spatial Pyramid Pooling Cross Stage Partial Combine (SPPCSPC) module. Subsequently, through a 1 × 1

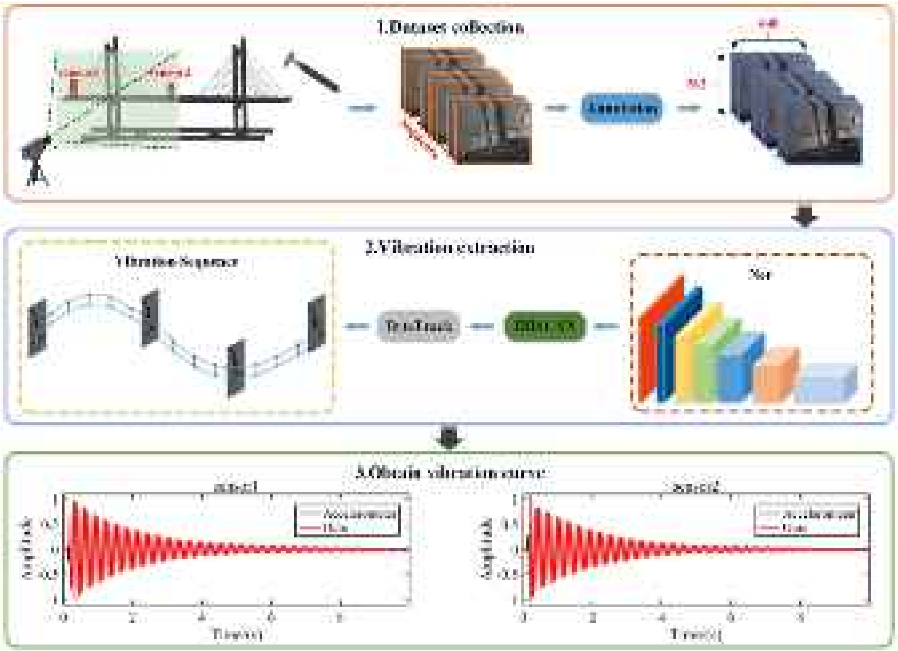

Fig. 2. Illustrate of the deep learning-based vibration displacement measurement algorithm.

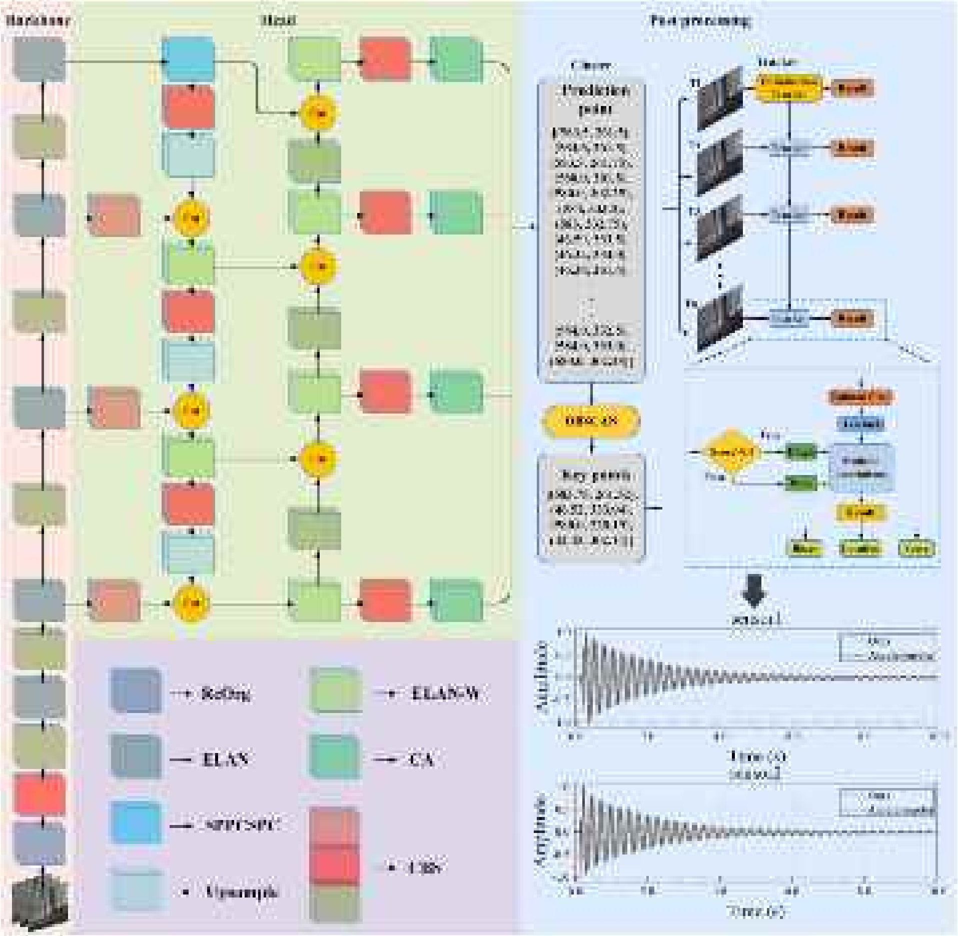

Fig. 3. Composition of the object detection algorithm framework in this study.

convolutional layer and upsampling, the resulting output is stacked with feature map P3, and then feature extraction is performed using ELAN-W to obtain feature map Q3. Repeat this process until Q1 is obtained. The bottom-up pathway enhancement section is similar to

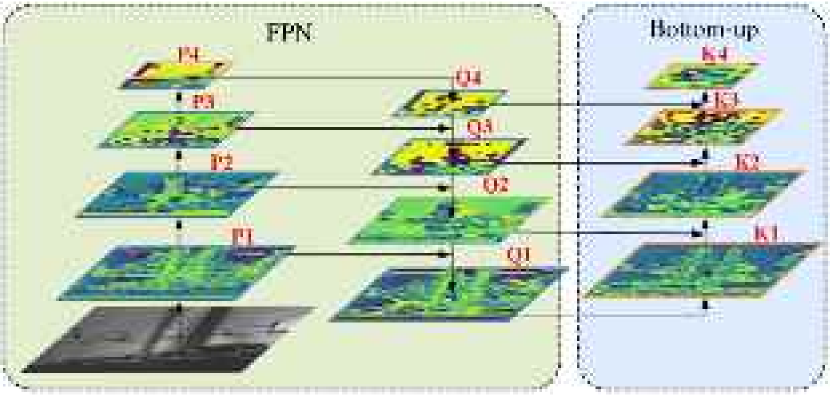

Fig. 4. PANet model architecture.

the top-down process in FPN. Feature map Q1 is processed through the ELAN-W module to produce feature map K1, which is further subsampled, stacked, and subjected to step-by-step feature extraction using a 3 × 3 convolution, ultimately resulting in K4.

To augment the network's ability to estimate the posture of flexible bridge structures and improve the precision of bridge keypoint localization, this study employs the Coordinate Attention (CA) mechanism, depicted in Fig. 5, to fortify the output feature layers of the Bottom-up pathway. This enhancement aims to extract more precise bridge vibration curves. Initially, the CA attention mechanism conducts global average pooling in both the width and height dimensions of the input feature maps, aggregating features along these spatial directions. After width-wise average pooling, we project the features onto a higher dimension, resulting in a feature layer of size $(C, H, 1)$. Following height-wise average pooling, the features are projected onto a wider dimension, yielding a feature layer of size $(C, 1, W)$. Subsequently, the width and height of the feature layers are transposed to a uniform dimension and stacked, followed by channel reduction through division by 'r'. Following convolution, normalization, and activation function processing, a feature layer with dimensions $(C/r, 1, (W, H))$ is acquired. Subsequently, by segregating and transposing the width and height, two distinct feature layers are derived: $(C, H, 1)$ and $(C, 1, W)$. Positional information along the height and width dimensions is extracted via 1 × 1 convolutions and sigmoid activation functions. Ultimately, the attention maps in both the width and height directions are multiplied with the input feature maps, augmenting the representation of positional information within the feature maps.

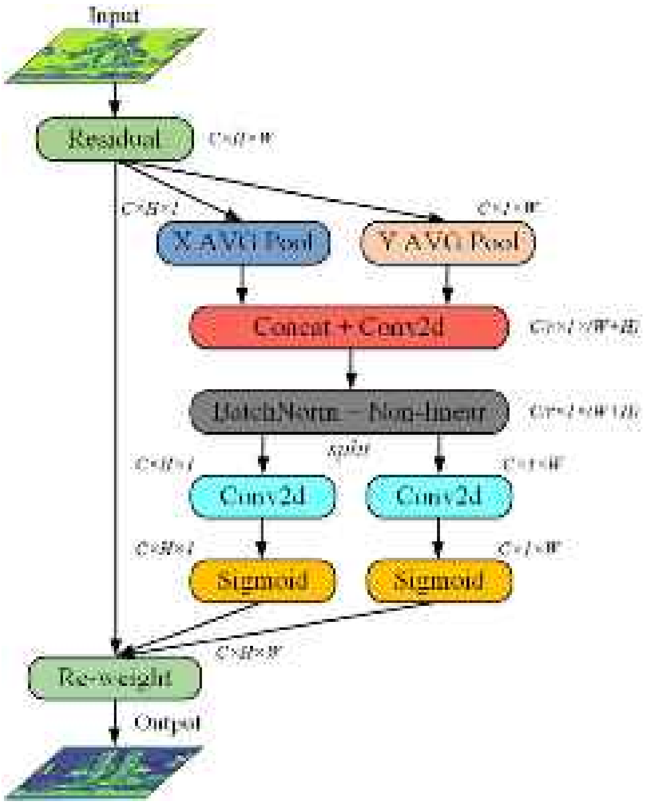

Fig. 5. Coordinate attention.

2.3. Loss

To enhance pose estimation precision, expedite loss convergence, and improve keypoint localization accuracy, we substitute the original network's $CIoU$ loss function with $SIoU$ in anchor box position regression. Unlike $CIoU$, $SIoU$ includes an angle loss, addressing the orientation disparity between predicted and ground-truth boxes.

The $SIoU$ loss function consists primarily of distance loss, angle loss, shape loss, and Intersection over Union (IoU) loss. The specific expressions are as follows: Δ denotes distance loss, Ω denotes shape loss, B denotes the predicted box, and B^{GT} denotes the ground-truth box.

$$SIoU = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{1}$$

$$IoU = \frac{\left|B \cap B^{GT}\right|}{\left|B \cup B^{GT}\right|} \tag{2}$$

The distance loss formula is given as follows: Here, Λ corresponds to the angle loss, c_w and c_h indicate the width and height of the predicted box's minimum enclosing rectangle, c_w represents the width of the minimum enclosing rectangle of the predicted box relative to the ground-truth box, while c_h corresponds to the height of the minimum enclosing rectangle of the predicted box relative to the ground-truth box. b_{c_x} and b_{c_y}respectively denote the width and height of the predicted box, while $b^{gt}_{c_x}$ and $b^{gt}_{c_y}$ represent the width and height of the ground-truth box.

$$\Delta = \sum_{t=x,y} (1 - e^{-\gamma \rho_t}) = 2 - e^{-\gamma \rho_x} - e^{-\gamma \rho_y} \tag{3}$$

$$\rho_x = \left(\frac{b^{gt}_{c_x} - b_{c_x}}{c_w}\right)^2 \tag{4}$$

$$\rho_y = \left(\frac{b^{gt}_{c_y} - b_{c_y}}{c_h}\right)^2 \tag{5}$$

$$\gamma = 2 - \Lambda \tag{6}$$

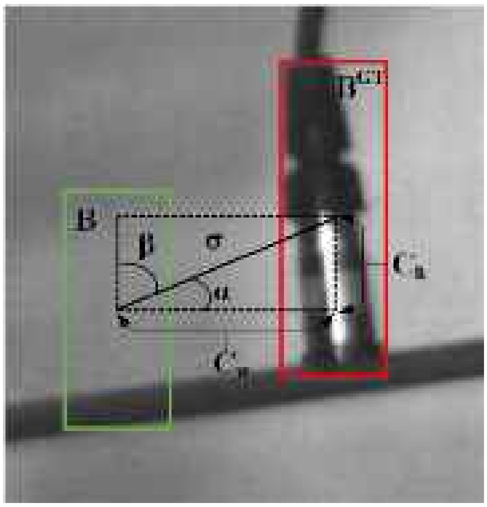

Fig. 6. SIoU angle loss.

Fig. 6 illustrates the schematic diagram of the angle loss, with C_w and C_h denoting the disparities in width and height between the predicted box and the ground-truth box's center points, while σ signifies the Euclidean distance between the center points of the predicted box and the ground-truth box. The angle loss calculation formula is as follow:

$$\Lambda = 1 - 2 \times \sin^2\left(\arcsin(x) - \frac{\pi}{4}\right) \tag{7}$$

$$x{=}\frac{C_h}{\sigma} = \sin(\alpha) \tag{8}$$

Object Keypoint Similarity (OKS) serves as a vital metric in the assessment of keypoints. When evaluating keypoints in the context of bridge targets, relying solely on Euclidean distance for measuring similarity proves inadequate. It is imperative to include scale information. In this paper, the algorithm predicts keypoints relative to anchor box centers and computes OKS individually for each keypoint. The final OKS loss or keypoint IOU loss is obtained by summing all individual keypoint OKS scores. Eq. (9) presents the specific expression for the OKS loss function:

$$L_{kpts}(s,i,j,k) = 1 - \sum_{n=1}^{N_{kpts}} OKS = 1 - \frac{\sum_{n=1}^{N_{kpts}} \exp\left(\frac{d_n^2}{2s^2k_n^2}\right)\delta\left(v_n > 0\right)}{\sum_{n=1}^{N_{kpts}} \delta\left(v_n > 0\right)} \tag{9}$$

In this context, d_n denotes the Euclidean distance between the predicted nth keypoint location and its actual position within the scene, K_n represents the keypoint's weight, s refers to the scale factor of the target object, and δ indicates the visibility status of each keypoint.

2.4. DBSCAN

In the original network's inference process, the non-maximum suppression algorithm is utilized to choose the keypoint with the highest confidence as the prediction. However, this method produces relatively inaccurate predictions and leads to stair-step patterns in the regressed vibration displacement of the target. Additionally, outdoor video acquisition is susceptible to factors like variations in lighting and complex environmental conditions, resulting in challenges such as image blurring and reduced visibility of target features. Consequently, keypoint with the highest confidence can exhibit substantial deviations from the actual keypoint, possibly even representing outliers. In the inference process, multiple keypoints with relatively high confidence are initially chosen as predicted data, and the DBSCAN clustering algorithm is utilized to determine the central positions of clustered target groups. Subsequently, the central points obtained after clustering replace the keypoints with the highest confidence as the predicted keypoints. Figs. 7 and 8 illustrate the positions of keypoints after clustering and the resulting vibration curves, respectively. It is noticeable that the original network's regressed vibration displacement curves demonstrate poor precision. Through the application of the DBSCAN clustering algorithm, the accuracy and robustness of target keypoint detection are enhanced, yielding smoother extracted vibration displacement curves that closely resemble the actual vibration displacement patterns.

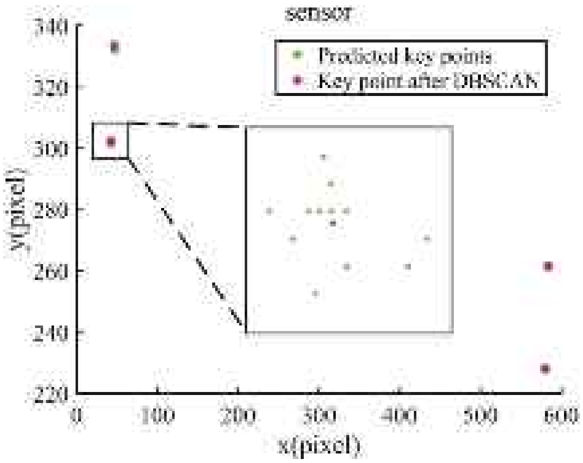

Fig. 7. Keypoints after DBSCAN clustering.

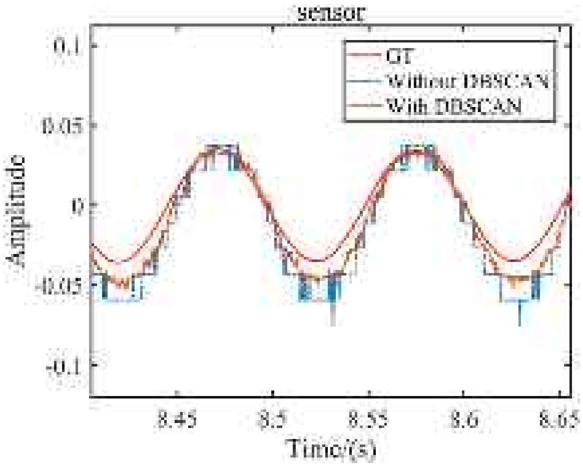

Fig. 8. The vibration curve extracted through the DBSCAN clustering algorithm is depicted.

2.5. Tracking section

Improving spatiotemporal coherence between frames in a video stream is beneficial for enhancing the smoothness of keypoints displacement curves in bridge target pose estimation. This paper introduces the ByteTrack multiple association tracking algorithm, enhancing high-precision keypoint detection. This integration strengthens the linkage between keypoint positions of consecutive frames in the video stream, leading to enhanced accuracy in vibration displacement measurements. Additionally, the inclusion of the tracking algorithm improves the method's robustness in challenging environmental conditions, including video noise and occlusions, rendering it suitable for real-world

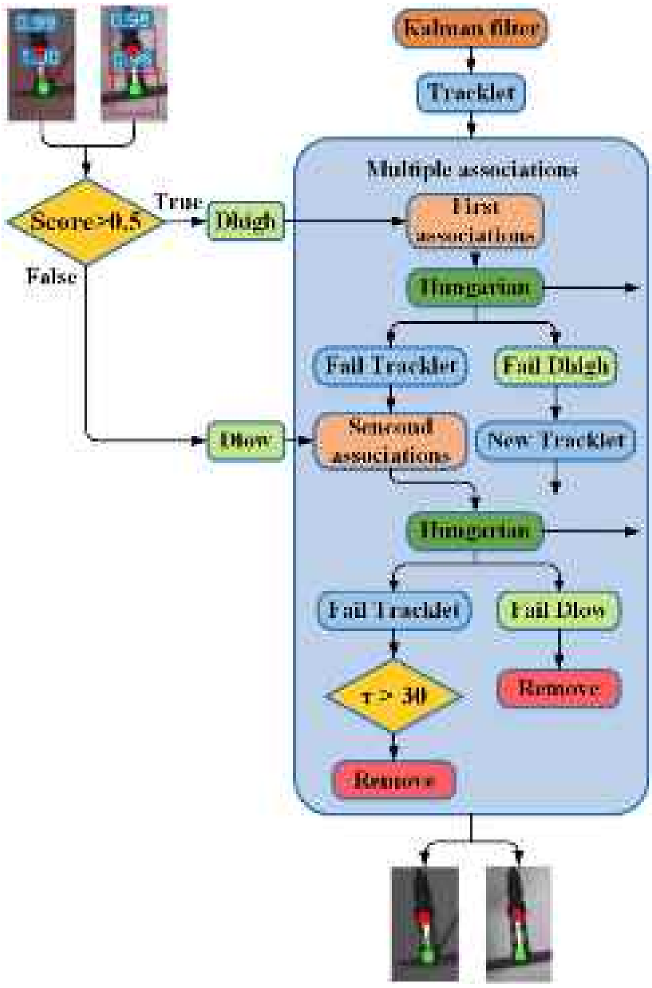

Fig. 9. Multi-association process diagram of the ByteTrack tracking algorithm.

outdoor bridge scenarios. Processed keypoint position information is linked to a 40 × 40 detection box for each keypoint after clustering. It serves as input to the tracking algorithm, along with keypoint confidence scores. The tracking algorithm utilizes Kalman filtering for position prediction within trajectories, employs the IoU metric to measure similarity between detection and prediction boxes, and ultimately uses the Hungarian algorithm for matching.

Fig. 9 illustrates the precise operational procedure of ByteTrack's multiple association. Initially, detection boxes are categorized into high-confidence and low-confidence groups, and Kalman filtering is employed to predict the new positions of trajectories in the current frame. Subsequently, in the first matching phase, high-confidence boxes are linked to pre-existing tracking trajectories. Similarity is assessed using the IoU metric between detection and prediction boxes, with matching performed through the Hungarian algorithm, which exploits this similarity measure. Subsequently, the Hungarian algorithm is employed for the second round of low-confidence box utilization, matching them with tracking trajectories not previously linked to high-confidence boxes in the initial phase. Low-confidence boxes failing to match in this secondary step are removed, and the trajectories that do not find a match are retained. Trajectories unpaired after 30 frames are subsequently excluded from further consideration. Lastly, unpaired high-confidence boxes are initialized as new trajectories.

2.6. Vibration extraction

During the image processing, the top-left point of the image sequence serves as the origin (0,0), while the bottom-right point represents the maximum value point (640,512) of the image. The method proposed in this study primarily detects the vertical displacement of bridges. After clustering and tracking, the predicted keypoints from the network retain the y-direction data of the prediction points. Subsequently, the y-direction data of the keypoints serve as the basis for determining the target vibration displacement. As illustrated in Fig. 10, the y-direction data of sensor 1 is 333, and that of sensor 2 is 228.19. By predicting a sequence of 20,000 frames of images, we obtained 20,000 sets of visual vibration data from sensors. It is noteworthy that the format of the data collected by the accelerometer sensor differs from that of the visual data. Therefore, we normalized the visual data and the vibration data obtained from the accelerometer sensor for comparison.

3. Experimental evaluation

In order to verify the feasibility and accuracy of this paper's algorithm for vibration displacement detection in large-span flexible structural bodies. In this section, the vibration of an outdoor large-span flexible body bridge is simulated by an excited cable-stayed bridge model, the acceleration sensor is placed at the position of the maximum vibration amplitude of the cable-stayed bridge model, and the feature points on the sensor are taken as the extraction points of the vibration of the cable-stayed bridge model, so as to construct the dataset. After training and validating various deep learning algorithms using this dataset, qualitative and quantitative comparisons are made between the visual displacement measurements obtained through different algorithms and the standard displacement offsets derived from the acceleration sensors.

3.1. Experimental setup and dataset preparation

In this section, we conducted experiments using an acceleration sensor (603C01 SNLW258034) in conjunction with a signal acquisition card (NI-9234), with a sampling frequency set at 25.6 kHz. These instruments were employed to capture the acceleration signals generated by the cable-stayed bridge model following impulsive hammer excitation. For high frame-rate image acquisition, we utilized high-speed industrial cameras (Qianyanlang 5F01) with a sampling image resolution set to 640 × 512 and a sampling frame rate of 2000 frames per second. In instances of insufficient natural lighting, supplementary illumination was provided using a Jinbei EF-200LED light source. All equipment was connected to a laptop (Honor MagicBook Pro) to synchronize time and sampling frequency. Fig. 11 illustrates the data collection platform for the model bridge vibration and the associated equipment.

The dataset of flexible bridge model data created in this paper encompasses a total of 20,000 continuous vibration frames, all acquired from the bridge model. In our experimental procedures, we downsampled the data obtained from the acceleration sensor to match the 2000 Hz sampling frequency of the camera. We use Labelme annotation software for 380 cable-stayed bridge model image data, to box the bridge target bounding box and label the target key points, and after labeling, we generate json files through Labelme, and use the transformation code to generate a format that supports training. Finally, the labeled dataset is divided into training set and validation set according to the ratio of 9:1 to construct a complete indoor bridge vibration dataset. All visual vibration measurement algorithms underwent training and validation using this identical dataset. The models were trained on an NVIDIA GeForce RTX 3090Ti with consistent settings, including 500 epochs and a batch size of 8, while retaining default parameters.

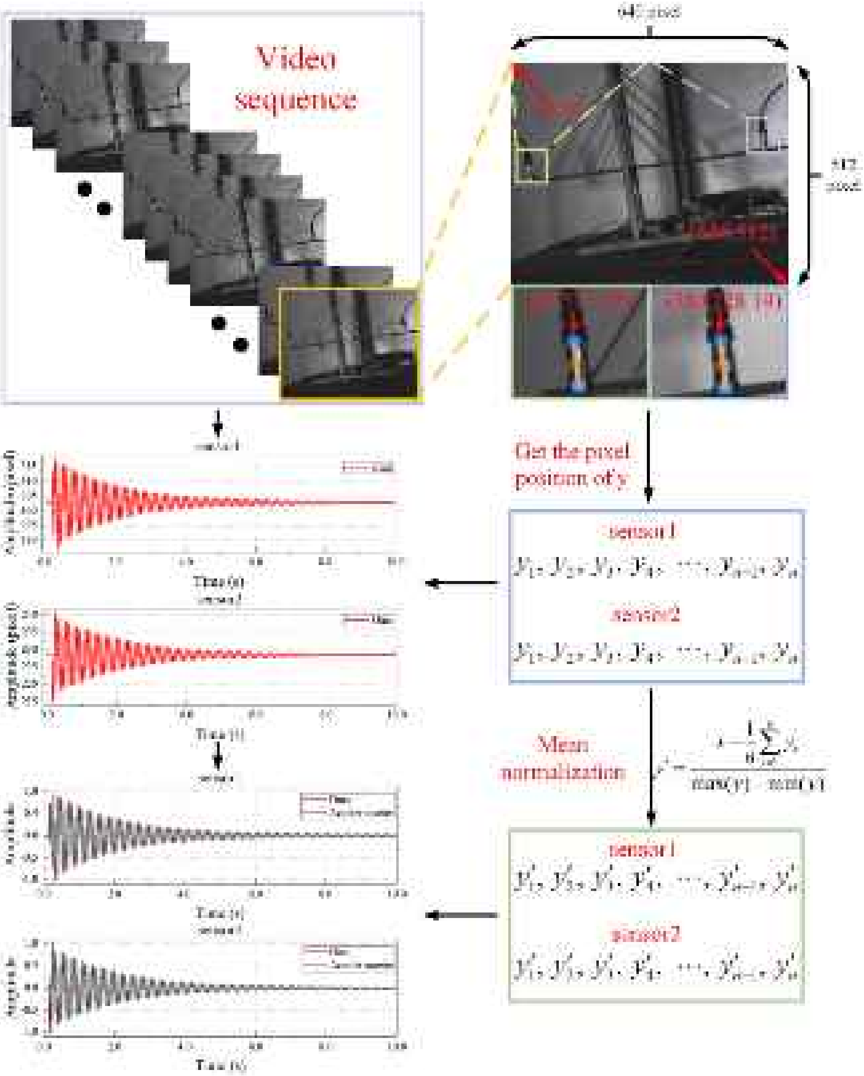
Video sequence
640 pixel
512 pixel
Get the pixel position of y
sensor1
$y_1, y_2, y_3, y_4, \cdots, y_{n-1}, y_n$
sensor2
$y_1, y_2, y_3, y_4, \cdots, y_{n-1}, y_n$
Mean normalization
sensor1
$y'_1, y'_2, y'_3, y'_4, \cdots, y'_{n-1}, y'_n$
sensor2
$y'_1, y'_2, y'_3, y'_4, \cdots, y'_{n-1}, y'_n$
Time (s)
Time (s)
Time (s)
Time (s)

Fig. 10. Vibration displacement extraction.

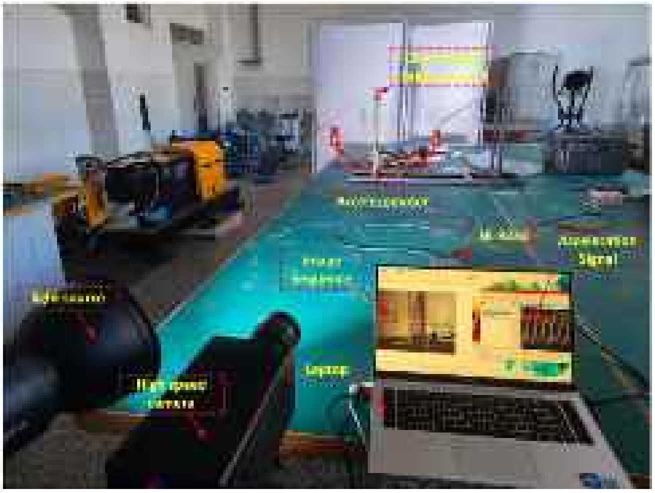

Fig. 11. Data acquisition equipment and acquisition platform.

3.2. Network model ablation study and evaluation metrics

In this section, we assess the enhancements introduced by the algorithm in this paper by employing various evaluation metrics. In order to showcase the effectiveness of these algorithmic improvements, we will conduct ablation experiments, and the corresponding detailed data is presented in Table 2. To demonstrate the accuracy of the vibration displacement measurements conducted by our algorithm, this study introduced Normalized Root Mean Square Error (NRMSE), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE) to assess the precision of the measurements. Quantitative comparisons were made by assessing the deviations between the values obtained from different visual measurement algorithms and those from the accelerometer sensor, aiming to evaluate the effects of various improvement methods on our algorithm.

The formulas for assessing measurements in the manuscript, including RMSE, NRMSE, MAE, and MAPE, are presented as follows:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(G_i - P_i)^2} \tag{10}$$

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(G_i - P_i)^2}}{G_{\max} - G_{\min}} \tag{11}$$

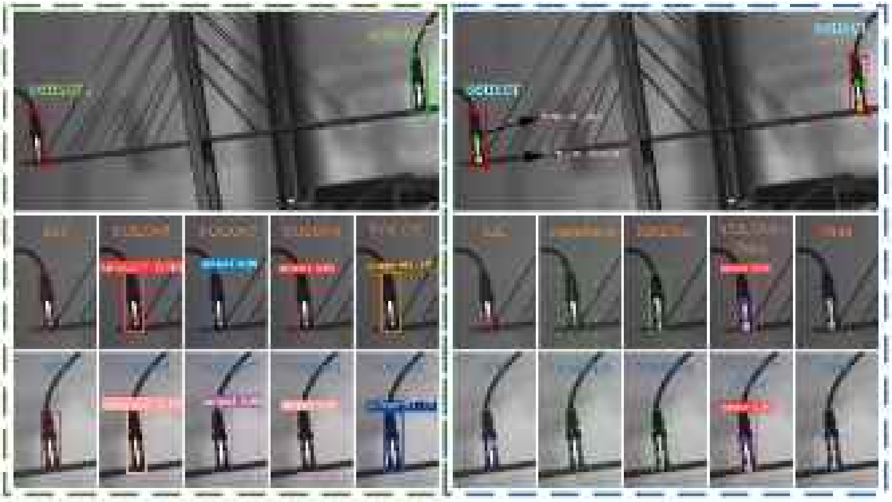

Fig. 12. Qualitative comparison of different methods.

Table 2
Ablation study.

NO.	Methods	NRMSE		mNRMSE	mMAE	mMAPE (%)	FPS
		sensor1	sensor2				
A	YOLOv7-Pose	0.0137	0.0254	0.0196	0.0308	1.5378	112.89
B	A+SIoU	0.0159	0.0185	0.0177	0.0223	1.1129	102.29
C	B+CA	0.0133	0.0212	0.0173	0.0209	1.0469	95.75
D	C+DBSCAN	0.0103	0.0169	0.0136	0.0105	0.5235	91.91
E	D+ByteTrack	0.0102	0.0168	0.0135	0.0104	0.5170	89.56

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|G_i - P_i\right| \tag{12}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{G_i - P_i}{G_i}\right| \tag{13}$$

where G_i represents the actual value of the target position in the ith frame, P_i represents the predicted value of the target position in the ith frame, $G_{\max}$ and $G_{\min}$ denote the maximum and minimum values of the actual values, respectively, and N represents the total number of images.

With the addition of the coordinate attention mechanism, the clustering algorithm and the tracking algorithm, the detection accuracy increases while the model inference time also increases. The introduction of the attention mechanism introduces additional parameters to learn the weights of the attention, which increases the weights obtained from the neural network training, which occupies higher computer memory and higher computational requirements, thus affecting the computational efficiency of the algorithm. Secondly, the clustering algorithm and tracking algorithm after the model inference will also increase the computational amount of the computer, resulting in a reduction of computational efficiency. The actual vibration frequency of large-span bridges is often within 10 Hz, and according to the Nyquist sampling theorem, the sampling frequency should not be less than twice the highest frequency in the analog signal spectrum, and the computational efficiency of the algorithm in this paper meets the requirements. Therefore, this method of improving detection accuracy by consuming a small amount of measurement time is worthwhile. From Table 2, it can be observed that each network ablation resulted in an improvement in network performance. The mNRMSE steadily decreased from 0.0196 to 0.0135, mMAE decreased from 0.0308 to 0.0104, and mMAPE decreased from 1.5378 to 0.5170. This process indicates that each improvement made in this study was effective.

3.3. Network model testing and analysis

We conducted a comprehensive comparative analysis of visual vibration detection using a range of algorithms, which included traditional template matching (referred to as MT), sparse optical flow method Lucas-Kanade, and dense optical flow method Farneback, object detection networks (YOLOv5, YOLOv7 [51], YOLOv8, YOLOX [52]), single-stage keypoint detection network (YOLOv8-Pose), and two-stage keypoint detection network (MMPose [53]). The outcomes obtained from these diverse algorithms were visually represented for comparative purposes. The template matching algorithm is limited in its ability as it can only accommodate a single template, thereby encountering substantial challenges when trying to match the correct position of the target template, especially when it undergoes deformation due to vibrations. In contrast, the YOLO series algorithms demonstrate their capabilities by regressing closely fitting bounding boxes, resulting in sub-pixel level precision for visual vibration measurements. Additionally, Fig. 12 provides a visual representation of the limitations of the Lucas-Kanade algorithm, which failed to maintain consistent positions for corner points and, consequently, cannot reliably track specified feature points. Furthermore, the two-stage pose detection algorithm, MMPose, exhibits a slower detection speed when compared to other algorithms.

This paper employs two accelerometer sensors depicted in Fig. 12 as the detection targets. The displacement signals obtained from these accelerometer sensors serve as the reference displacement curve, hereafter referred to as the Ground Truth (GT) curve. This GT curve is then juxtaposed with positioning information predicted by different algorithms, enabling a qualitative analysis of the disparities among these algorithms. The corresponding results of the time-domain signal comparisons are illustrated in Figs. 13 and 14. Due to the nature of the template matching algorithm, which involves sliding a template within the target image and assessing the similarity between the template image and the target region, the detection accuracy achieved is at the pixel-level precision. This results in the extraction of a vibration curve exhibiting periodic extreme value oscillations. The extracted vibration displacement curve by YOLOv5 exhibits a downward offset phenomenon, with discontinuities observed under low-amplitude vibration conditions. YOLOv7's extracted vibration displacement curve also experiences a downward offset under low-amplitude vibrations, resulting in a stepped curve. YOLOv8 similarly exhibits a downward offset under low-amplitude vibrations when extracting data from sensor1 as the target, and a stepping phenomenon is observed when sensor2 is the target. In the case of YOLOX, when extracting data from sensor1 as the target, the vibration displacement curve not only displays a

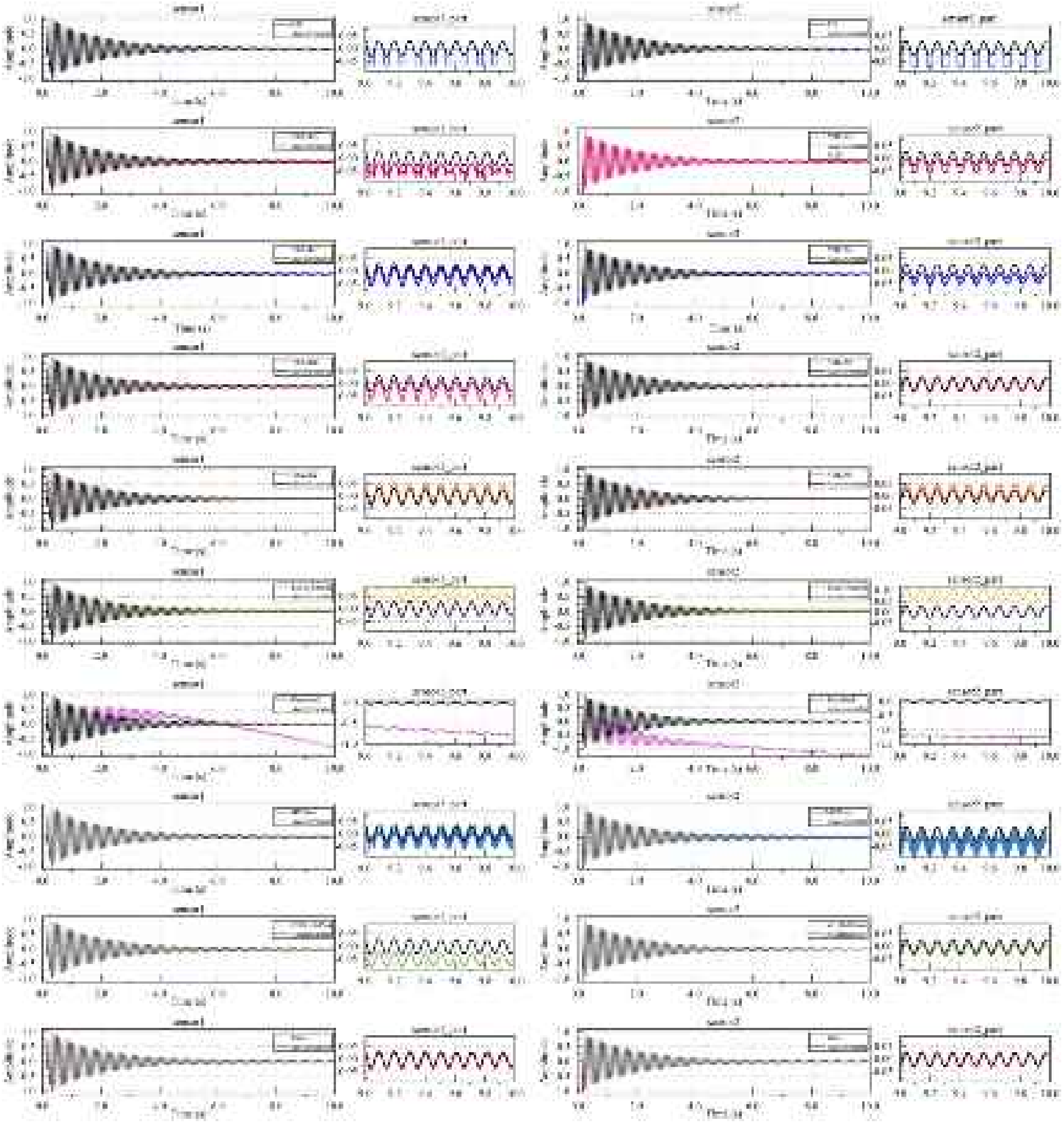

Fig. 13. Time-domain comparison of target image displacement signals on the bridge model with acquisition frequency of 2000 Hz.

downward offset but also exhibits an upward shift in the 1.3–2.5 s interval. However, its fitting performance is less satisfactory under low-amplitude vibrations when sensor2 is the target.

Lucas-Kanade has a higher detection speed compared to the algorithm in this article, and the extracted vibration displacement information matches the vibration displacement relatively accurately in the 0–2 s time period. However, as we delve into the later sections characterized by low-frequency vibrations, an interesting phenomenon emerges—a noticeable upward displacement shift in comparison to the standard vibration curve algorithm. The Farneback algorithm, by detecting the optical flow information at fixed positions in the video, was unable to track targets. Even though accurate frequency information could be obtained, the algorithm exhibited poor performance in capturing the actual vibration displacement information of the targets, resulting in significant deviations. MMPose, being a two-stage algorithm, heavily relies on the precision of detection box positions extracted using Faster-RCNN [54] and the accuracy of heatmap keypoint detection. This reliance results in a displacement curve that exhibits an overall downward shift, making it a less natural fit for the GT curve. Additionally, when sensor2 becomes the target, the extracted vibration displacement curve experiences a downward shift under the influence of low-frequency vibrations. When measuring vibrations using YOLOv8-Pose with two accelerometer sensors as targets, the extracted vibration displacement curves exhibit varying degrees of downward offset under low-frequency vibrations. In stark contrast, the algorithm we introduce in this study demonstrates its capability by extracting vibration displacement curves that remain free from offset or step phenomena, irrespective of whether the vibrations are high or low in amplitude. It becomes increasingly apparent that while comparative algorithms manage to capture displacement fluctuations akin to the GT curve, they grapple with stability and accuracy issues, rendering them less effective in matching accelerometer-derived vibration displacement curves.

When compared to its algorithmic counterparts, our proposed algorithm distinguishes itself in terms of both target detection precision

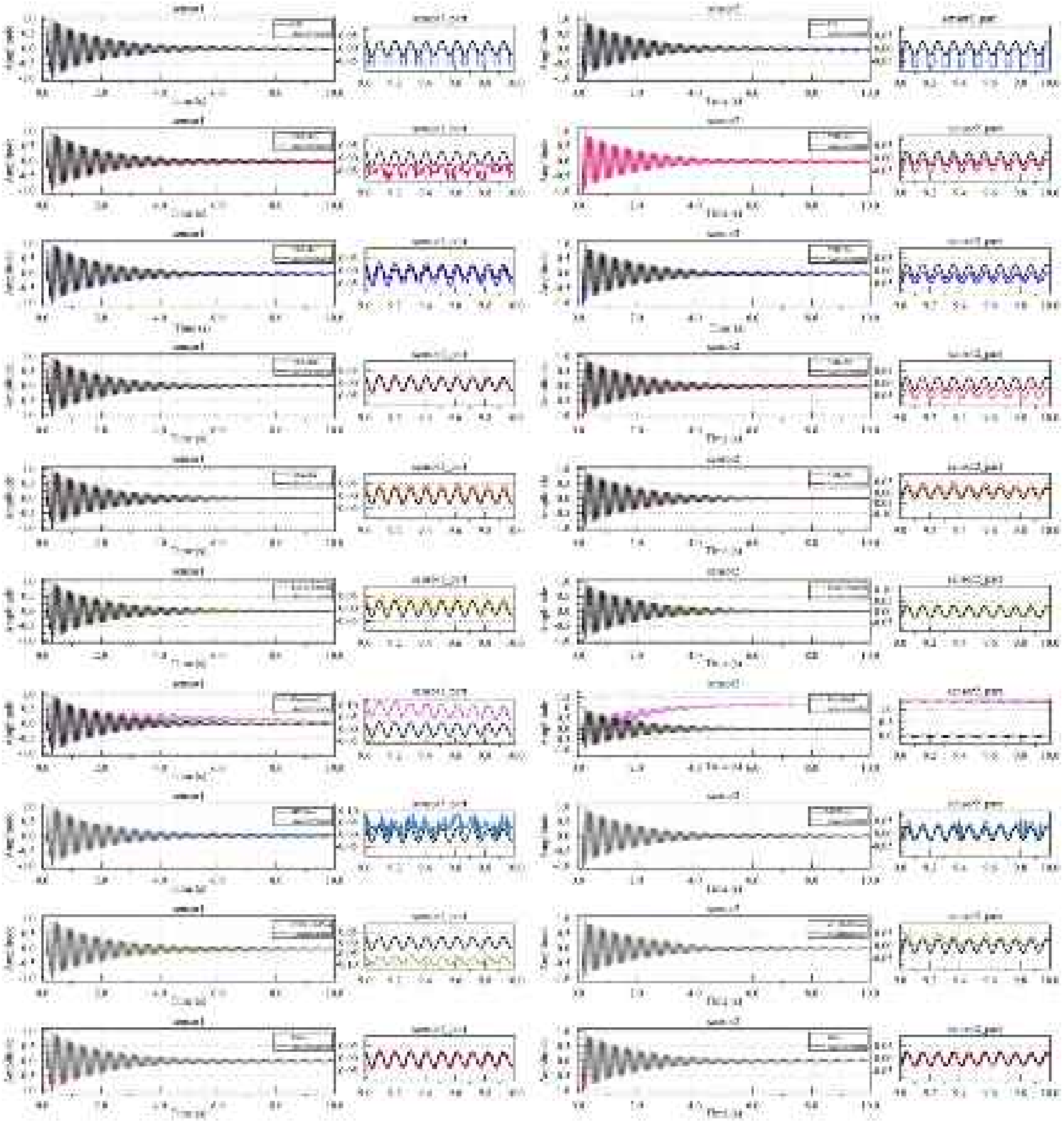

Fig. 14. Time-domain comparison of the target image displacement signal on the bridge model with acquisition frequency of 1000 Hz.

and stability. Consequently, the extracted vibration displacement information aligns more closely with the curves obtained from accelerometer data. This affirmation underscores the practical potential of our algorithm in the realm of bridge monitoring applications.

To investigate the impact of different sampling frequencies on vibration extraction algorithms, this study sampled 20,000 frames of images and downsampled them to obtain a vibration video of 1000 Hz, as illustrated in Fig. 14. It was observed that lower-frequency visual vibration extraction has a relatively minor effect on visual vibration measurement. The vibration displacement curve obtained by the algorithm in this study exhibited a higher degree of fit with the vibration displacement curve obtained by the accelerometer.

The quantitative comparison results of different visual measurement algorithms are presented in Tables 3 and 4. By comparing the NRMSE, MAE, and MAPE, it was observed that the error in the vibration displacement curve extracted by this study was the smallest compared to that obtained by the accelerometer sensor. The detection speed was found to be sufficient to meet the requirements of practical bridge vibration monitoring. Qualitative and quantitative assessments further bolster the credibility of our proposed algorithm, highlighting its attributes of high precision and robustness. These findings underscore its suitability for the demanding task of large-span bridge vibration detection applications.

4. Analysis of real bridge structural outdoor scenarios

In outdoor settings, the fluctuations in lighting and weather conditions introduce unpredictable noise into the captured images, presenting a substantial hurdle in the realm of visual vibration measurement. To evaluate the resilience of our algorithm and establish a comparative analysis with other benchmark algorithms under outdoor conditions for vibration measurement on large-span suspension bridges, this study conducted experimental validation in outdoor settings. More precisely, our approach underwent verification on both the Humen Bridge, which encounters deformations resulting from vortex-induced vibrations, and the Longjiang Bridge, subject to vibrations stemming from routine vehicular traffic during its normal operation.

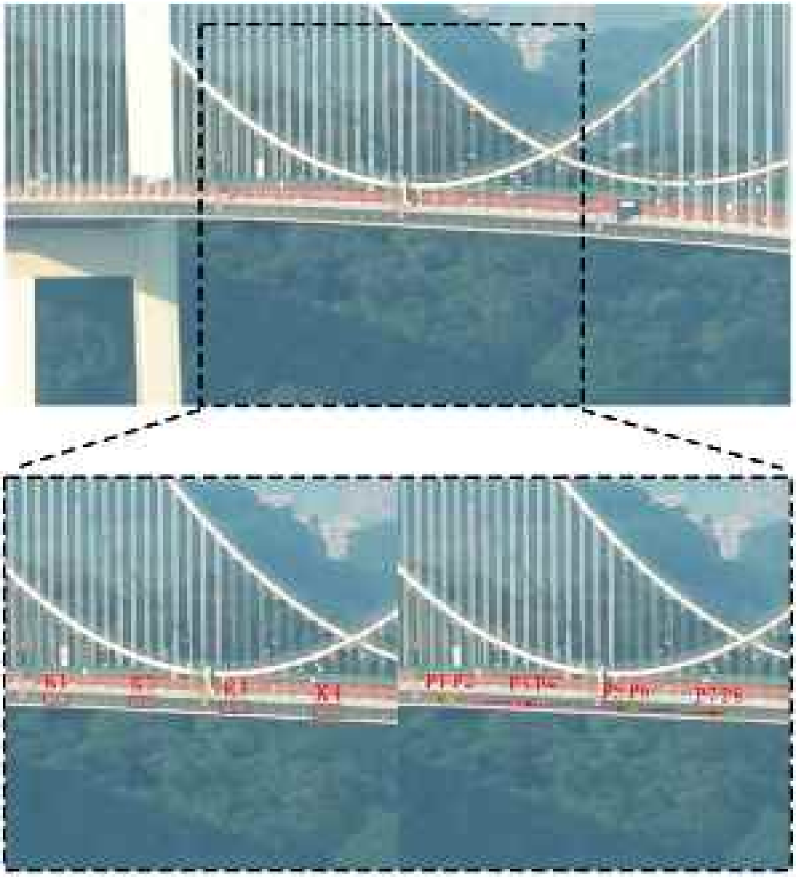

Fig. 15. Data collection at the Humen Bridge.

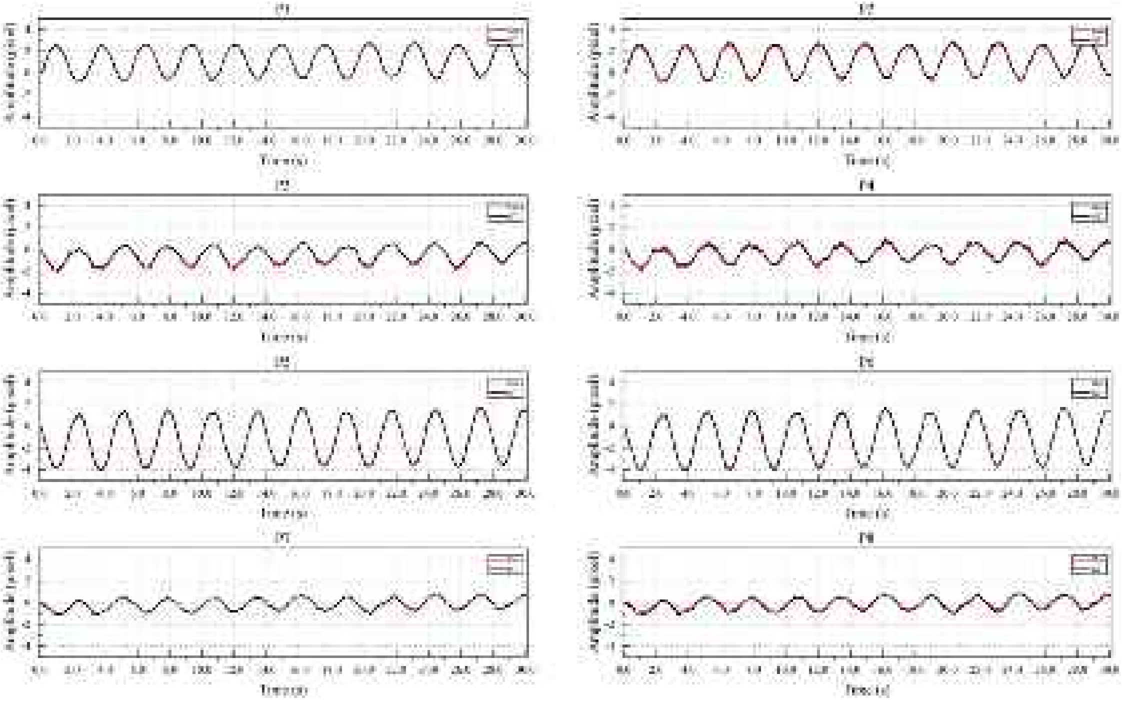

Fig. 16. Vibration displacement curve for keypoint detection of the Humen Bridge.

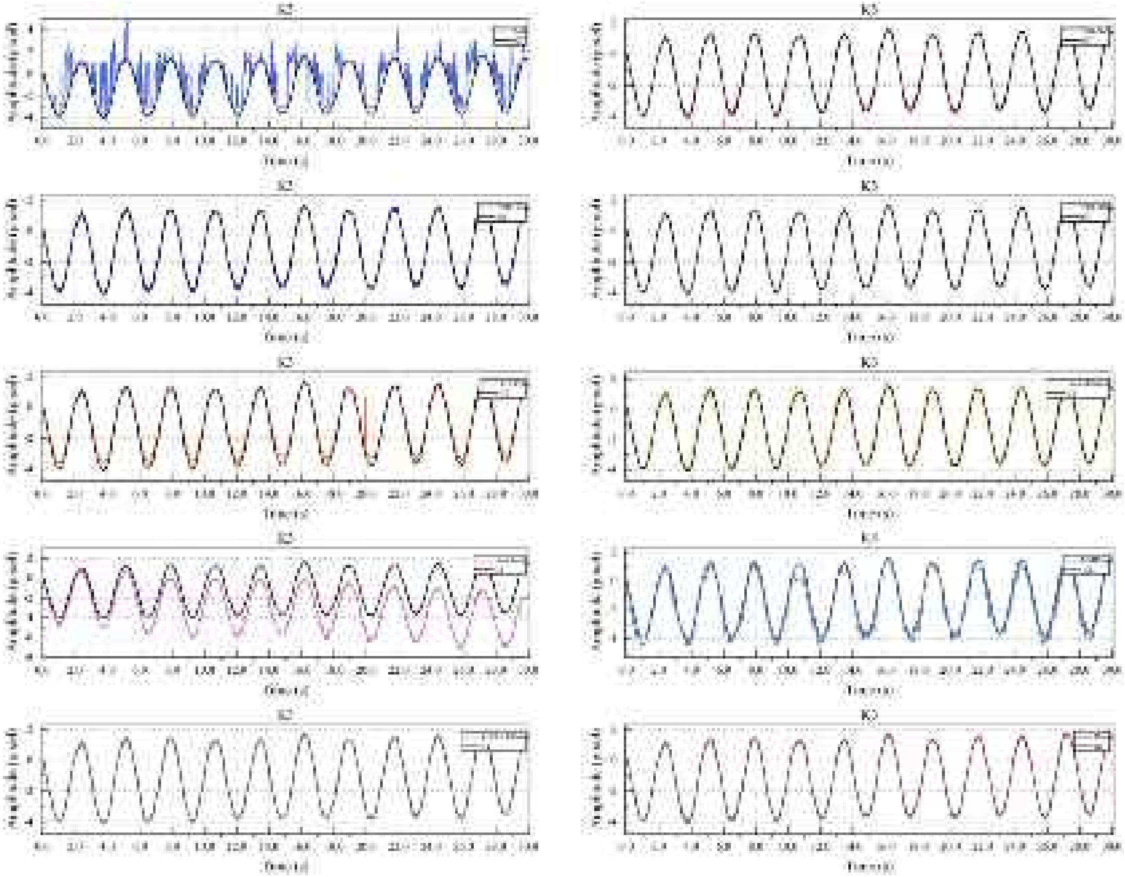

Fig. 17. Vibration displacement curve of the Humen Bridge.

Fig. 18. The Longjiang Bridge data collected using smartphones.

4.1. Validation of the Humen Bridge

4.1.1. Datasets and configuration

On May 5, 2020, maintenance work necessitated the addition of a temporary barrier wall, standing at a height of 1.2 m, to the Guangzhou Humen Bridge. This modification disrupted the bridge's original streamlined aerodynamic shape, leading to vortex-induced vibrations at low wind speeds, ultimately resulting in the closure of the bridge passage.

To conduct experiments, video samples were obtained from the Humen Bridge under the influence of vortex-induced vibrations, utilizing a sampling frequency of 30 FPS. From the original videos (as depicted in Fig. 15), images of dimensions 640 × 640 pixels were extracted to construct the dataset. In order to capture points with different amplitudes as comprehensively as possible and to measure the vibration positions of the bridge as extensively as possible, we selected K1, K2, K3, K4 as the targets for the object detection algorithm, and P1, P2, P3, P4, P5, P6, P7, P8 as the target points for keypoint detection.

In this study, 900 video frames were meticulously annotated from the collected image sequence. The annotated positional data was employed to regress the vibration curves, which served as GT curves. Subsequently, 380 selected images were divided into training and testing sets using a 9:1 ratio.

This study encompassed both displacement qualitative analysis and numerical quantitative comparisons between the vibration curves extracted through different algorithms within the pixel coordinate system

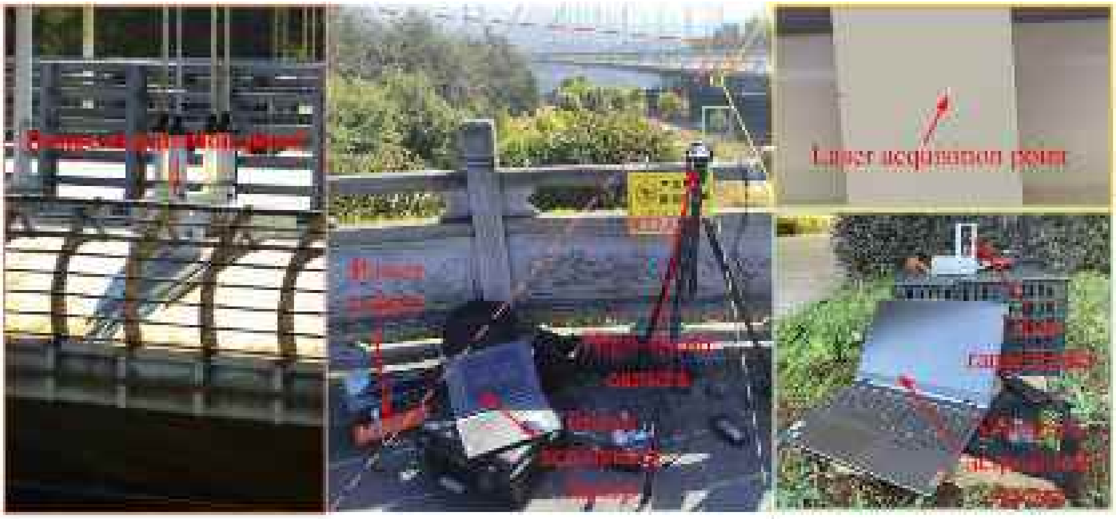

Fig. 19. The Longjiang Bridge data collected using industrial camera.

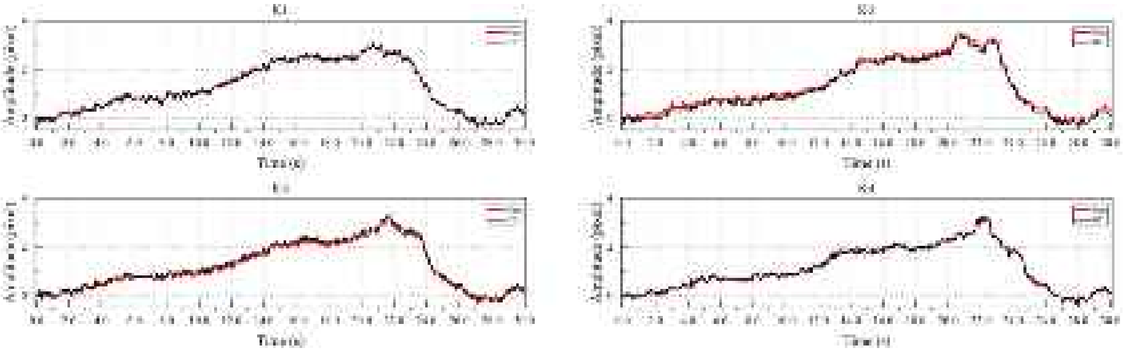

Fig. 20. Vibration displacement curve of keypoint detection for the Longjiang Bridge.

Table 3
Quantitative comparison between different visual displacement detection algorithms at a collection frequency of 2000 Hz.

Algorithms	NRMSE		mNRMSE	mMAE	mMAPE (%)	FPS
	sensor1	sensor2				
MT	0.0167	0.0268	0.0217	0.0361	1.5806	72.89
YOLOv5	0.0207	0.0221	0.0214	0.0328	1.6388	119.21
YOLOv7	0.0129	0.0231	0.0180	0.0252	1.2574	178.90
YOLOv8	0.0155	0.0310	0.0232	0.0200	0.9987	92.59
YOLOX	0.0134	0.0188	0.0161	0.0191	0.9554	175.60
Lucas-Kanade	0.0254	0.0349	0.0366	0.0540	2.6966	537.59
Farneback	0.1635	0.4497	0.6132	0.5584	27.9195	14.73
MMpose	0.0129	0.0279	0.0204	0.0298	1.4899	16.10
YOLOv8-Pose	0.0225	0.0227	0.0226	0.0319	1.5915	172.41
Ours	0.0102	0.0168	0.0135	0.0103	0.5170	89.56

Table 4
Quantitative comparison between different visual displacement detection algorithms at a collection frequency of 1000 Hz.

Algorithms	NRMSE		mNRMSE	mMAE	mMAPE (%)	FPS
	sensor1	sensor2				
MT	0.0170	0.0273	0.0222	0.0324	1.6181	73.64
YOLOv5	0.0225	0.0220	0.0223	0.0342	1.7122	121.32
YOLOv7	0.0148	0.0238	0.0193	0.0276	1.3773	175.44
YOLOv8	0.0118	0.0276	0.0197	0.0270	1.3475	69.93
YOLOX	0.0133	0.0229	0.0181	0.0207	1.0367	157.83
Lucas-Kanade	0.0157	0.0205	0.0181	0.0243	1.2130	549.32
Farneback	0.0666	0.5059	0.2863	0.5359	26.7938	13.89
MMpose	0.0247	0.0228	0.0238	0.0357	1.7833	15.80
YOLOv8-Pose	0.0352	0.0246	0.0299	0.0486	2.4276	171.36
Ours	0.0106	0.0174	0.0140	0.0119	0.5920	88.42

and the GT curves. All vibration measurement algorithms underwent uniform training and validation within the same dataset and deep learning environment.

4.1.2. Experimental validation and result analysis

Due to vortex-induced vibrations, the Humen Bridge experiences mechanical transverse wave vibrations. In this study, we selected eight target keypoints for conducting multi-point measurements of the bridge's vibrations. Consequently, we acquired eight sets of vibration displacement data from distinct positions on the bridge, which we subsequently subjected to qualitative comparisons with the GT curve. Fig. 16 illustrates the comparative results. Upon analyzing Fig. 16, it becomes apparent that adjacent keypoints exhibit subtle variations in vibration, characterized by minimal differences in both amplitude and waveform. In contrast, keypoints that are spaced farther apart, specifically P1, P3, P5, and P7, display more pronounced discrepancies in wave peaks and waveforms. Of particular note, P5 and P6 are positioned in close proximity to the mechanical wave crest, resulting in the highest vibration amplitude, whereas P7 and P8 are situated near the mechanical wave nodes, leading to the smallest vibration amplitude. These findings provide insight into the correspondence between the vibrations at different bridge characteristic points and the overall bridge vibrations.

In order to further validate the performance of the proposed algorithm for monitoring vibration displacement on actual bridges, this

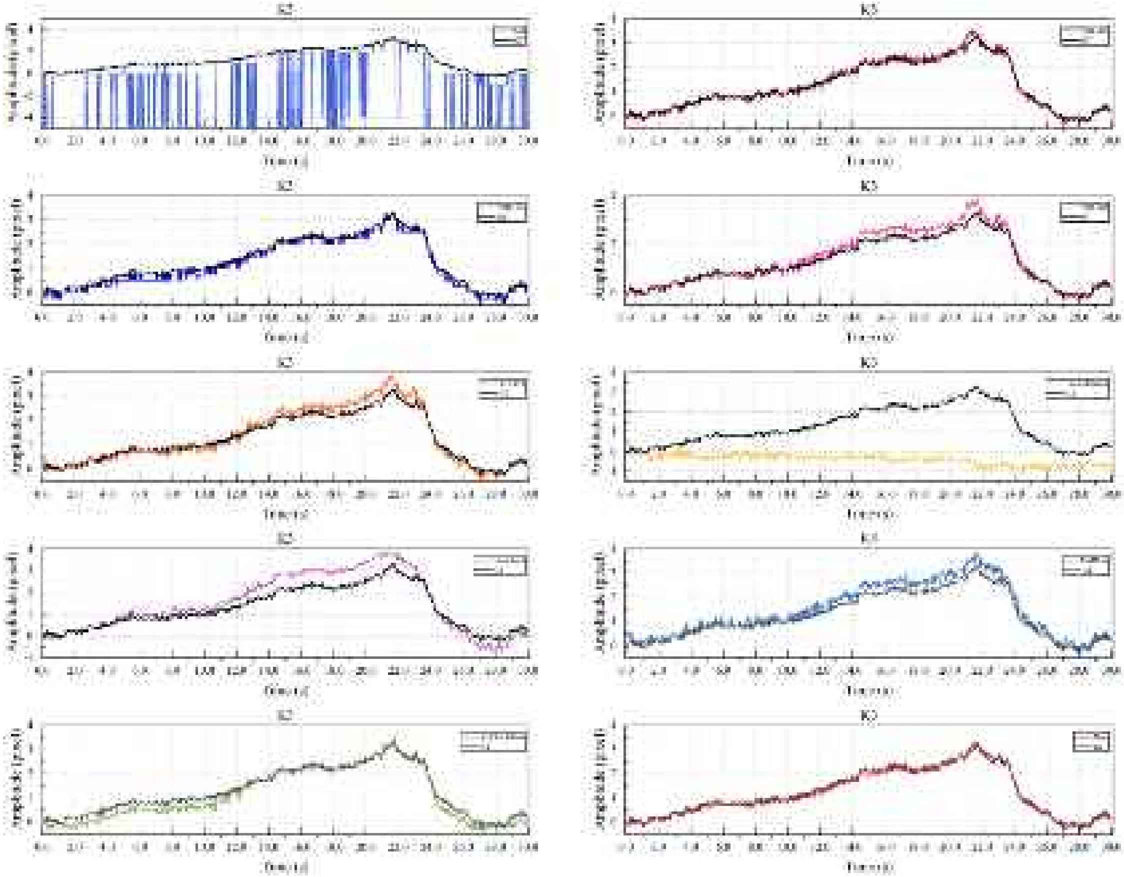

Fig. 21. Image vibration displacement curve captured by smartphone of Longjiang Bridge.

Table 5
Quantitative comparison among various visual displacement detection algorithms for the Humen Bridge.

Algorithms	RMSE(K1)	RMSE(K2)	RMSE(K3)	RMSE(K4)	mRMSE	mMAE	mMAPE (%)	FPS
MT	6.3955	6.7799	1.7775	6.4774	5.3575	3.9325	80.3064	73.69
YOLOv5	0.1170	0.1633	0.1080	0.0711	0.1149	0.0954	2.0497	71.42
YOLOv7	0.1904	0.1564	0.1281	0.1724	0.1618	0.1346	2.8149	159.70
YOLOv8	0.1419	0.1449	0.0815	0.1371	0.1263	0.1064	2.2087	67.73
YOLOX	0.1595	0.1545	0.3046	0.1667	0.1963	0.1602	3.5598	156.13
Lucas-Kanade	0.2637	0.4733	0.2795	1.1798	0.5490	0.4654	9.8946	850.08
Farneback	0.9115	1.3365	2.0318	0.3656	1.1614	1.0148	23.2156	9.37
MMpose	0.4129	0.3586	0.4306	0.4030	0.4012	0.3269	6.9821	15.60
YOLOv8-Pose	0.0930	0.1943	0.1266	0.1065	0.1301	0.1075	2.3562	66.23
Ours	0.0507	0.1167	0.0811	0.0462	0.0736	0.0618	1.3822	86.58

study selects two specific targets: K3, situated closest to the crest of the vibration wave, and P5 as the primary vibration detection points on the bridge. Various algorithms are employed to assess their performance using data from the Humen Bridge dataset. Subsequently, the extracted vibration displacement curves are compared with the GT curve, illustrated in Fig. 17. It becomes apparent that the MT algorithm yields results akin to those obtained in laboratory conditions, achieving only pixel-level accuracy. However, it struggles to precisely detect minor target vibrations, resulting in relatively substantial errors when compared to other algorithms. Except for the MT algorithm, the majority of algorithms successfully capture the vibration displacement of the detection targets on the bridge. To quantitatively evaluate the accuracy and stability of these algorithms, this study calculates the RMSE, MAE and MAPE between the vibration displacement curves obtained by the various algorithms and the GT curve. As indicated in Table 5, our proposed algorithm demonstrates the closest alignment with the GT curve.

4.2. Experimental validation of the Longjiang Bridge

4.2.1. Datasets and configuration

The Longjiang Bridge, located in Yunnan Province, China, spans the Longjiang River, with impressive dimensions – a total length of 2,470.58 m, a bridge deck width of 33.5 m, and a towering height of 129.703 m, making it Asia's largest suspension bridge in terms of span. To further validate the generalizability of the algorithm proposed in this study, visual vibration measurements were conducted on the Longjiang Bridge using two types of sensors: smartphones and industrial cameras. The smartphone recorded videos for 30 s, while the industrial camera recorded for 300 s. Both sensors captured videos at a frame rate of 30 FPS, and the processed images were resized to 640 × 640 pixels. In the smartphone dataset, the protective casing at the connection between the suspension cable and the stiffening girder was used as the detection target, while in the industrial camera dataset, the connection point between the suspension cable and the stiffening girder served as the detection target. Field data collection is illustrated in Figs. 18 and 19. The dataset's construction followed

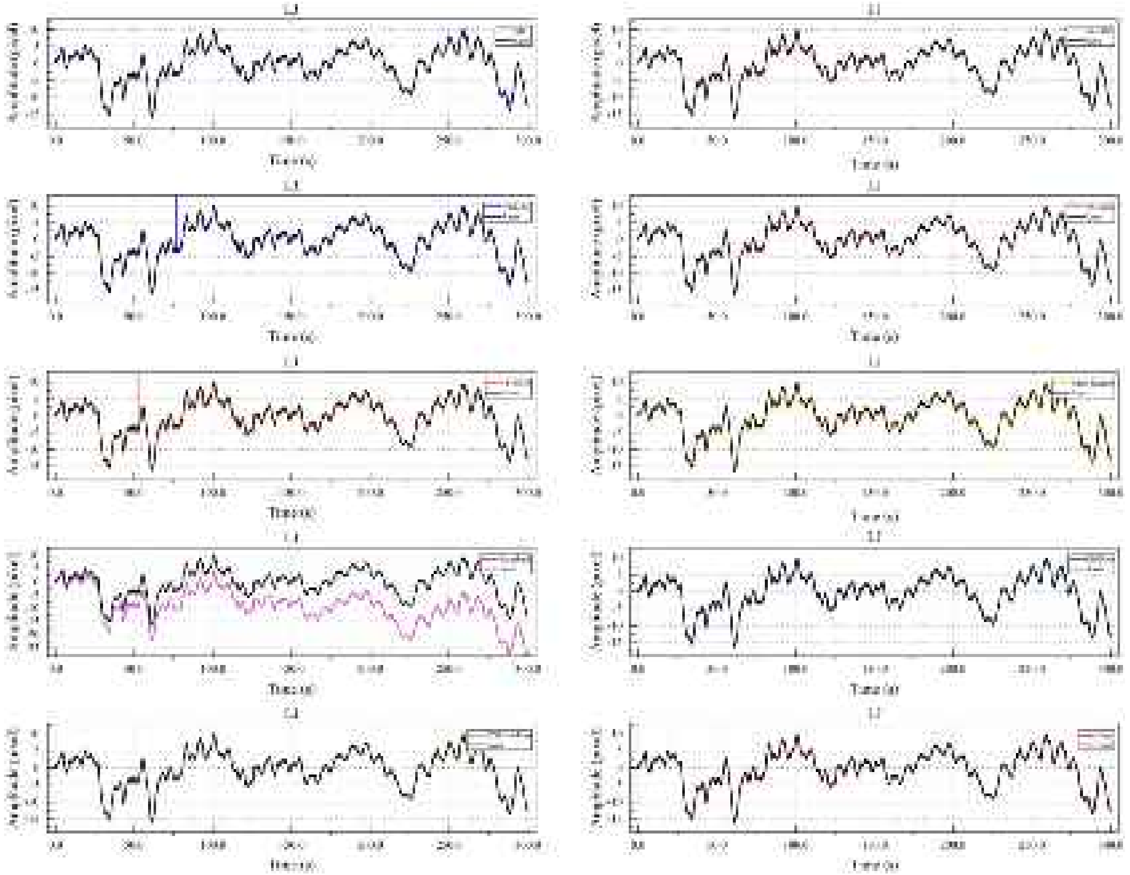

Fig. 22. Image vibration displacement curve captured by industrial camera of Longjiang Bridge.

the methods outlined in Section 4.1 regarding quantity, selection, and sample distribution. Qualitative displacement analysis and quantitative comparisons were conducted on displacement curves extracted from various algorithms and the GT curve in pixel coordinates to verify the algorithm’s efficacy. All vibration measurement algorithms underwent uniform training and validation within the same dataset and deep learning environment.

4.2.2. Experimental validation and result analysis

We utilized the algorithm presented in this paper to detect targets at various locations, allowing us to generate vibration displacement curves corresponding to different target positions. From Fig. 20, it can be observed that the vibration waveform at different positions is generally similar, and it exhibits a good fit with the GT curve. This finding highlights the algorithm’s stability in measuring vibrations across diverse target positions, consistently delivering effective vibration displacement measurements.

To further validate the performance of our algorithm in real-world bridge vibration displacement monitoring, we selected targets K1, K2, K3, and K4 as the objectives for vibration detection on the bridge. We employed various algorithms to measure the vibrations of these targets and compared the displacement signals of target K3 generated by different detection algorithms with the GT curve, as depicted in Fig. 21. Fig. 21 highlights that complex outdoor lighting conditions have a significant impact on the accuracy of the template matching algorithm. It shows in a misidentification of K2 as K3 during the detection process, leading to a substantial displacement in the detected position. The Lucas-Kanade algorithm’s successful operation relies on the constant brightness of feature points in the video stream. However, in the complex bridge environment with continuous vehicle traffic, this prerequisite is disrupted, preventing the algorithm from obtaining stable feature points and impeding its ability to track the target reliably.

In comparison to conventional algorithms, deep learning algorithms consistently detect targets, although there are subtle differences in detection performance among various deep learning algorithms. From Fig. 22, it can be observed that when using an industrial camera to capture bridge vibrations, both the YOLOv7 and YOLOX algorithms exhibit discontinuities in vibration target extraction. The Farneback algorithm gradually shifts the vibration curve downward as the extraction time increases. As shown in Tables 6 and 7, compared to other deep learning algorithms, the algorithm proposed in this study demonstrates higher accuracy and stability, and it exhibits good vibration extraction performance under different sensors.

5. Conclusion

This article proposes a bridge vibration displacement extraction method based on keypoint detection and verifies it through experiments on a cable-stayed bridge model in a laboratory environment, as well as on the outdoor Humen Bridge and Longjiang Bridge. The accuracy of bridge displacement measurement is improved by introducing an attention mechanism and modifying the loss function in the original network. Additionally, the DBSCAN clustering and ByteTrack tracking of key points enhance the stability of keypoint vibration extraction. In the cable-stayed bridge model, the NRMSE improved from 0.0196 to 0.0135, the MAE improved from 0.0308 to 0.0104, and the MAPE improved from 1.5378% to 0.5170%. The algorithm presented in this paper achieved an RMSE of 0.0736, an MAE of 0.0618, and a MAPE of 1.3822% under the Humen Bridge dataset. For the Longjiang Bridge dataset captured with a cell phone, the RMSE was 0.0813, the MAE was 0.0714, and the MAPE was 3.2263%. For the Longjiang Bridge dataset captured with an industrial camera, the RMSE was 0.3852, the MAE was 0.2990, and the MAPE was 1.6982%. These results verify

Table 6
Quantitative comparison of different visual displacement detection algorithms for the Longjiang Bridge using smartphones.

Algorithms	RMSE(K1)	RMSE(K2)	RMSE(K3)	RMSE(K4)	mRMSE	mMAE	mMAPE (%)	FPS
MT	0.5833	0.5150	4.8884	0.2699	1.5641	1.1346	50.6178	67.52
YOLOv5	0.2405	0.1439	0.1066	0.2385	0.1823	0.1402	6.3400	83.34
YOLOv7	0.1378	0.1894	0.1656	0.1577	0.1626	0.1305	5.8976	154.30
YOLOv8	0.1573	0.2747	0.2260	0.2633	0.2303	0.1898	8.5951	68.49
YOLOX	0.1417	0.3071	0.2267	0.1177	0.1983	0.1502	6.7871	162.31
Lucas-Kanade	0.4284	1.9900	1.9423	7.9752	3.0839	2.1880	99.6496	733.86
Farneback	2.2496	1.5546	0.4614	1.0495	1.3288	1.1527	51.9827	8.06
MMpose	0.2013	0.4031	0.2776	0.2210	0.2757	0.2390	10.8119	15.70
YOLOv8-Pose	0.1386	0.2023	0.2288	0.2239	0.1984	0.1523	6.8839	68.49
Ours	0.0461	0.1288	0.1030	0.0475	0.0813	0.0714	3.2263	88.31

Table 7
Quantitative comparison of different visual displacement detection algorithms for the Longjiang Bridge using industrial cameras.

Algorithms	mRMSE	mMAE	mMAPE (%)	FPS
MT	0.4958	0.3883	2.2053	24.22
YOLOv5	0.3904	0.3069	1.7430	85.56
YOLOv7	2.9376	0.3586	2.0366	58.82
YOLOv8	0.3936	0.3112	1.7675	52.63
YOLOX	4.2467	0.5451	3.0960	151.34
Lucas-Kanade	0.8917	0.8188	4.6500	702.02
Farneback	9.6341	8.7572	49.7347	14.14
MMpose	0.4237	0.3423	1.9441	17.60
YOLOv8-Pose	0.3949	0.3079	1.7486	67.11
Ours	0.3852	0.2990	1.6982	55.56

the universality and superiority of the algorithm. After qualitative and quantitative comparisons with various deep learning-based algorithms, it was found that the keypoint detection algorithm proposed in this paper has high measurement accuracy and robustness in the visual vibration measurement of large-span bridges.

We found that traditional visual vibration algorithms have a simpler reasoning approach compared to the algorithm proposed in this paper. Lucas Kanade has a faster inference speed and demonstrates good detection accuracy in laboratory environments. Some deep learning algorithms, such as YOLOv5 and YOLOv7, also have fast inference speeds. We will continue to improve our algorithms, combine the advantages of other algorithms, enhance their detection speed and accuracy, and build a lightweight network architecture. We plan to use the TensorRT acceleration model to speed up target image inference and deploy this algorithm on the Nvidia Jetson Orin NX device to build an edge bridge vibration measurement system.

CRediT authorship contribution statement

Ruiyang Sun: Writing – original draft, Software, Methodology, Conceptualization. **Sen Wang:** Visualization, Software, Methodology, Data curation. **Mao Li:** Validation, Software, Methodology. **Yang Zhu:** Methodology, Data curation.

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

The authors do not have permission to share data.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 52065035, and in part by the General Program of Basic Research Projects of Yunnan Provincial Department of Science and Technology under Grant 202301AT070468.

References

[1] Y. Fujino, D. Siringoringo, Vibration mechanisms and controls of long-span bridges: a review, Struct. Eng. Int. 23 (3) (2013) 248–268.

[2] F. Ni, J. Zhang, M.N. Noori, Deep learning for data anomaly detection and data compression of a long-span suspension bridge, Comput.-Aided Civ. Infrastruct. Eng. 35 (7) (2020) 685–700.

[3] Z. Fang, W. Wang, Y. Cao, Q. Li, Y. Lin, T. Li, D. Wu, S. Wu, Reciprocating compressors intelligent fault diagnosis under multiple operating conditions based on adaptive variable scale morphological filter, Measurement 224 (2024) 113778.

[4] X. Jia, L. He, H. Zhang, Effect of turbine rotor disc vibration on hot gas ingestion and rotor-stator cavity flow, Aerosp. Sci. Technol. 98 (2020) 105719.

[5] K. Feng, Y. Xiao, Z. Li, Z. Jiang, F. Gu, Gas turbine blade fracturing fault diagnosis based on broadband casing vibration, Measurement 214 (2023) 112718.

[6] C. Kralovec, M. Schagerl, Review of structural health monitoring methods regarding a multi-sensor approach for damage assessment of metal and composite structures, Sensors 20 (3) (2020) 826.

[7] Z. Ma, J. Choi, H. Sohn, Continuous bridge displacement estimation using millimeter-wave radar, strain gauge and accelerometer, Mech. Syst. Signal Process. 197 (2023) 110408.

[8] F. Geng, Z. Bai, H. Zhang, Y. Yao, C. Liu, P. Wang, X. Chen, L. Du, X. Li, B. Han, et al., Contactless and continuous blood pressure measurement according to caPTT obtained from millimeter wave radar, Measurement 218 (2023) 113151.

[9] G. Zhang, W. Zhao, J. Zhang, Bridge distributed stiffness identification of continuous beam bridge based on microwave interferometric radar technology and rotation influence line, Measurement 220 (2023) 113353.

[10] S. Häberling, M. Rothacher, Y. Zhang, J.F. Clinton, A. Geiger, Assessment of high-rate GPS using a single-axis shake table, J. Geod. 89 (2015) 697–709.

[11] S.B. Im, S. Hurlebaus, Y.J. Kang, Summary review of GPS technology for structural health monitoring, J. Struct. Eng. 139 (10) (2013) 1653–1664.

[12] J.G. Chen, A. Davis, N. Wadhwa, F. Durand, W.T. Freeman, O. Büyüköztürk, Video camera–based vibration measurement for civil infrastructure applications, J. Infrastruct. Syst. 23 (3) (2017) B4016013.

[13] S.J. Rothberg, M. Allen, P. Castellini, D. Di Maio, J. Dirckx, D. Ewins, B.J. Halkon, P. Muyshondt, N. Paone, T. Ryan, et al., An international review of laser Doppler vibrometry: Making light work of vibration measurement, Opt. Lasers Eng. 99 (2017) 11–22.

[14] P.-J. Chun, T. Yamane, Y. Maemura, A deep learning-based image captioning method to automatically generate comprehensive explanations of bridge damage, Comput.-Aided Civ. Infrastruct. Eng. 37 (11) (2022) 1387–1401.

[15] H. Yoon, H. Elanwar, H. Choi, M. Golparvar-Fard, B.F. Spencer Jr., Target-free approach for vision-based structural system identification using consumer-grade cameras, Struct. Control Health Monit. 23 (12) (2016) 1405–1416.

[16] T. Khuc, F.N. Catbas, Completely contactless structural health monitoring of real-life structures using cameras and computer vision, Struct. Control Health Monit. 24 (1) (2017) e1852.

[17] J.G. Chen, N. Wadhwa, Y.-J. Cha, F. Durand, W.T. Freeman, O. Buyukozturk, Modal identification of simple structures with high-speed video using motion magnification, J. Sound Vib. 345 (2015) 58–71.

[18] X. Zhang, W. Wang, K. Chen, W. Li, D. Zhang, L. Tian, Five dimensional movement measurement method for rotating blade based on blade tip timing measuring point position tracking, Mech. Syst. Signal Process. 161 (2021) 107898.

[19] Y. Wang, W. Hu, J. Teng, Y. Xia, Full-field displacement measurement of long-span bridges using one camera and robust self-adaptive complex pyramid, Mech. Syst. Signal Process. 215 (2024) 111451.

[20] Z. Shang, Z. Shen, Multi-point vibration measurement and mode magnification of civil structures using video-based motion processing, Autom. Constr. 93 (2018) 231–240.

[21] J. Zhang, Q. Zhang, T. Jiang, C. Hou, Vibration displacement measurement method based on vision Gaussian fitting and edge optimisation for rotating shafts, Measurement 232 (2024) 114699.

[22] R. Song, C. Ren, J. Cheng, C. Li, X. Yang, Non-contact human respiratory rate measurement based on two-level fusions of video and FMCW radar information, Measurement 222 (2023) 113604.

[23] S. Liu, L. Yu, W. Niu, J. Wang, Z. Zhong, J. Huang, M. Shan, Fast and accurate visual vibration measurement via derivative-enhanced phase-based optical flow, Mech. Syst. Signal Process. 209 (2024) 111089.

[24] T. Yang, J. Wu, Z. Yu, X. Meng, L.-Q. Chen, Tracking and measurement of periodic rotating blades after video recombination based on center mark recognition, Measurement 229 (2024) 114412.

[25] C.-Z. Dong, O. Celik, F.N. Catbas, Marker-free monitoring of the grandstand structures and modal identification using computer vision methods, Struct. Health Monit. 18 (5–6) (2019) 1491–1509.

[26] S. Bhowmick, S. Nagarajaiah, Spatiotemporal compressive sensing of full-field Lagrangian continuous displacement response from optical flow of edge: Identification of full-field dynamic modes, Mech. Syst. Signal Process. 164 (2022) 108232.

[27] L. Luo, M.Q. Feng, Z.Y. Wu, Robust vision sensor for multi-point displacement monitoring of bridges in the field, Eng. Struct. 163 (2018) 255–266.

[28] L. Luo, M.Q. Feng, J. Wu, L. Bi, Modeling and detection of heat haze in computer vision based displacement measurement, Measurement 182 (2021) 109772.

[29] R. Del Sal, L. Dal Bo, E. Turco, A. Fusiello, A. Zanarini, R. Rinaldo, P. Gardonio, Structural vibration measurement with multiple synchronous cameras, Mech. Syst. Signal Process. 157 (2021) 107742.

[30] Y. Yang, C. Dorn, T. Mancini, Z. Talken, G. Kenyon, C. Farrar, D. Mascareñas, Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification, Mech. Syst. Signal Process. 85 (2017) 567–590.

[31] D. Feng, M.Q. Feng, Vision-based multipoint displacement measurement for structural health monitoring, Struct. Control Health Monit. 23 (5) (2016) 876–890.

[32] C. Dong, X. Ye, T. Jin, Identification of structural dynamic characteristics based on machine vision technology, Measurement 126 (2018) 405–416.

[33] P. Cheng, W.-J. Li, W.-L. Chen, D.-L. Gao, Y. Xu, H. Li, Computer vision-based recognition of rainwater rivulet morphology evolution during rain–wind-induced vibration of a 3D aeroelastic stay cable, J. Wind Eng. Ind. Aerodyn. 172 (2018) 367–378.

[34] M. Debella-Gilo, A. Kääb, Sub-pixel precision image matching for measuring surface displacements on mass movements using normalized cross-correlation, Remote Sens. Environ. 115 (1) (2011) 130–142.

[35] J. Zhao, Y. Bao, Z. Guan, W. Zuo, J. Li, H. Li, Video-based multiscale identification approach for tower vibration of a cable-stayed bridge model under earthquake ground motions, Struct. Control Health Monit. 26 (3) (2019) e2314.

[36] J. Guo, et al., Dynamic displacement measurement of large-scale structures based on the Lucas–Kanade template tracking algorithm, Mech. Syst. Signal Process. 66 (2016) 425–436.

[37] J.M.W. Brownjohn, Y. Xu, D. Hester, Vision-based bridge deformation monitoring, Front. Built Environ. 3 (2017) 23.

[38] J. Guo, X. Wu, J. Liu, T. Wei, X. Yang, X. Yang, B. He, W. Zhang, Non-contact vibration sensor using deep learning and image processing, Measurement 183 (2021) 109823.

[39] S. Lin, S. Wang, T. Liu, X. Liu, C. Liu, Accurate measurement of bridge vibration displacement via deep convolutional neural network, IEEE Trans. Instrum. Meas. 72 (2023) 5020016.

[40] R. Cao, Y. Zhao, Y. Gao, X. Huang, L. Zhang, Effects of flow rates and layer thicknesses for aggregate conveying process on the prediction accuracy of aggregate gradation by image segmentation based on machine vision, Constr. Build. Mater. 222 (2019) 566–578.

[41] R. Wang, J. Li, S. An, H. Hao, W. Liu, L. Li, et al., Densely connected convolutional networks for vibration based structural damage identification, Eng. Struct. 245 (2021) 112871.

[42] Y. Zhang, P. Liu, X. Zhao, Structural displacement monitoring based on mask regions with convolutional neural network, Constr. Build. Mater. 267 (2021) 120923.

[43] Y. Shao, L. Li, J. Li, S. An, H. Hao, Computer vision based target-free 3D vibration displacement measurement of structures, Eng. Struct. 246 (2021) 113040.

[44] M. Li, S. Wang, T. Liu, X. Liu, C. Liu, Rotating box multi-objective visual tracking algorithm for vibration displacement measurement of large-span flexible bridges, Mech. Syst. Signal Process. 200 (2023) 110595.

[45] R. Yang, S. Wang, X. Wu, T. Liu, X. Liu, Using lightweight convolutional neural network to track vibration displacement in rotating body video, Mech. Syst. Signal Process. 177 (2022) 109137.

[46] Q. Hou, D. Zhou, J. Feng, Coordinate attention for efficient mobile network design, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 13713–13722.

[47] Z. Gevorgyan, SIoU loss: More powerful learning for bounding box regression, 2022, arXiv preprint arXiv:2205.12740.

[48] Y. Zhang, P. Sun, Y. Jiang, D. Yu, F. Weng, Z. Yuan, P. Luo, W. Liu, X. Wang, Bytetrack: Multi-object tracking by associating every detection box, in: European Conference on Computer Vision, Springer, 2022, pp. 1–21.

[49] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, Yolov4: Optimal speed and accuracy of object detection, 2020, arXiv preprint arXiv:2004.10934.

[50] S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759–8768.

[51] C.-Y. Wang, A. Bochkovskiy, H.-Y.M. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7464–7475.

[52] Z. Ge, S. Liu, F. Wang, Z. Li, J. Sun, Yolox: Exceeding yolo series in 2021, 2021, arXiv preprint arXiv:2107.08430.

[53] A. Sengupta, F. Jin, R. Zhang, S. Cao, Mm-pose: Real-time human skeletal posture estimation using mmwave radars and CNNs, IEEE Sens. J. 20 (17) (2020) 10032–10044.

[54] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, Adv. Neural Inf. Process. Syst. 28 (2015).