Received: 12 February 2023 | Revised: 23 June 2023 | Accepted: 27 July 2023

DOI: 10.1002/msd2.12083

RESEARCH ARTICLE

IJMSD International Journal of Mechanical System Dynamics Open Access WILEY

Nodes2STRNet for structural dense displacement recognition by deformable mesh model and motion representation

Jin Zhao[1,2,3,4] | Hui Li[1,2,3] | Yang Xu[1,2,3]

[1]Key Lab of Smart Prevention and Mitigation of Civil Engineering Disasters of the Ministry of Industry and Information Technology, Harbin Institute of Technology, Harbin, China

[2]Key Lab of Structures Dynamics Behavior and Control of the Ministry of Education, Harbin Institute of Technology, Harbin, China

[3]School of Civil Engineering, Harbin Institute of Technology, Harbin, China

[4]Department of System Design and Simulation, Goldwind Science and Technology Co., Ltd., Beijing, China

Correspondence

Assoc. Prof. Yang Xu, School of Civil Engineering, Harbin Institute of Technology, 150090 Harbin, China.
Email: xyce@hit.edu.cn

Funding information

National Natural Science Foundations of China, Grant/Award Numbers: 52192661, 51921006, 52008138; China Postdoctoral Science Foundations, Grant/Award Numbers: BX20190102, 2019M661286; Heilongjiang Provincial Natural Science Foundation, Grant/Award Number: LH2022E070; Heilongjiang Province Postdoctoral Science Foundations, Grant/Award Numbers: LBH-TZ2016, LBH-Z19064

Abstract

Displacement is a critical indicator for mechanical systems and civil structures. Conventional vision-based displacement recognition methods mainly focus on the sparse identification of limited measurement points, and the motion representation of an entire structure is very challenging. This study proposes a novel Nodes2STRNet for structural dense displacement recognition using a handful of structural control nodes based on a deformable structural three-dimensional mesh model, which consists of control node estimation subnetwork (NodesEstimate) and pose parameter recognition subnetwork (Nodes2PoseNet). NodesEstimate calculates the dense optical flow field based on FlowNet 2.0 and generates structural control node coordinates. Nodes2PoseNet uses structural control node coordinates as input and regresses structural pose parameters by a multilayer perceptron. A self-supervised learning strategy is designed with a mean square error loss and L_2 regularization to train Nodes2PoseNet. The effectiveness and accuracy of dense displacement recognition and robustness to light condition variations are validated by seismic shaking table tests of a four-story-building model. Comparative studies with image-segmentation-based Structure-PoseNet show that the proposed Nodes2STRNet can achieve higher accuracy and better robustness against light condition variations. In addition, NodesEstimate does not require retraining when faced with new scenarios, and Nodes2PoseNet has high self-supervised training efficiency with only a few control nodes instead of fully supervised pixel-level segmentation.

KEYWORDS

structural dense displacement recognition, deformable structural mesh model, deep-learning-based monocular vision, self-supervised learning

1 | INTRODUCTION

The identification of displacement of objects is significant for mechanical systems and civil structures. A number of displacement measurement techniques have been proposed, such as linear variable differential transformer (LVDT), global positioning system (GPS), laser displacement transducers, and so forth. However, the LVDT and laser displacement transducer are inconvenient for field monitoring of full-scale systems and structures, and the resolution and sampling rate of GPS are limited.

Vision-based displacement measurement methods have been proposed for several decades.[1–9] At the early stage, the methods required installing artificial targets on structures. Wahbeh et al.[10] installed light-emitting diode lights on bridges to measure structural displacement in low-light conditions, such as at night. However, the artificial targets are either labor-consuming or challenging to access in field applications. Measurement methods based on the visual features of structures without artificial targets have been proposed, for example, feature point matching,[11,12] digital image correlation (DIC),[13] correlation filter, and motion amplification. The feature point matching algorithm includes scale-invariant feature transform (SIFT),[14,15] speeded-up robust features,[16] and Kanade–Lucas–Tomasi (KLT).[17] The algorithms calculate the image feature points through the feature operator and match the feature points on different images. In addition to the manual design of the feature operator, Dong and Catbas[18] designed a deep-learning-based feature operator (Visual Geometry Group) and combined it with SIFT to identify the displacement of a two-span bridge model. The correlation filter method[19–21] used a self-training iterative algorithm to track the initial object. Zhao et al.[22] combined support correlation filters and KLT to improve the robustness of displacement measurement. However, all these feature point matching methods can only obtain the single point displacement of the structure. Helfrick et al.[23] and Baqersad et al.[24] applied the DIC algorithm to measure the vibration and rotation of different types of structures and obtain the displacement field. Almeida et al.[25] used the DIC algorithm based on a set of images measuring the planar deformation. In addition to the fixed camera, unmanned aerial vehicles (UAVs) are also widely used to obtain videos and identify structural displacement.[26,27] The laser-embedded light detection and ranging technology is implemented into UAVs to scan three-dimensional (3D) point clouds to determine structural displacement.[28,29]

In the above vision-based methods, the identification procedure for displacement can be divided into two steps: identification of the pixel displacement on the image, and then conversion of the pixel displacement into real-world displacement using scaling factor or camera matrix transformation. Structure-PoseNet[30] proposed a structural displacement identification method based on a deformable structural 3D mesh model (DSMM), and the displacement is directly obtained from the coordinates in the mesh model. Generally, 3D dynamic displacement recognition of a structure includes the identification of 3D models and model poses. To identify the structural pose parameters, image features should be first refined from video frames. These features should contain the motion of structural components, exclude image background, structural texture, light illumination, and shadow, and be sensitive to subtle motion. Deep-learning-based computer vision techniques can be used to extract structural features. To realize the abovementioned goals, an appropriate two-dimensional (2D) representation of structural motion should be adopted. The semantic segmentation mask was selected as the structural motion representation in Structure-PoseNet.[30] However, it is sensitive to the variation of light conditions, and the semantic segmentation mask lacks gradient variations, limiting its effectiveness in identifying dense displacement. Moreover, the training efficiency is limited because two new subnetworks of ParaNet and CompNet need to be retrained when the resolution of the input image changes.

In addition to semantic segmentation, dense optical flow[31,32] can extract structural motion features in the video. Dense optical flow recognizes the motion velocity field of pixels in the image while ignoring other unnecessary information. Optical flow is sensitive to small pixel movements, which is significant for motion identification. After the representations of structural motion features are obtained, they are further converted into structural pose parameters. This process can be accomplished through deep-learning networks.

In this study, Nodes2StrNet is proposed for dense structural displacement identification, which consists of a control node estimation subnetwork (NodesEstimate) and a pose parameter recognition subnetwork (Nodes2PoseNet). The NodesEstimate subnetwork takes each video frame as the input and outputs the 2D position of control nodes, and the Nodes2PoseNet subnetwork takes the coordinates of control nodes as the input and outputs the structural pose parameters. Finally, the dense displacement of the structure is obtained based on the deformed structural 3D mesh model.

The remainder of this paper is organized as follows. Section 2 introduces the proposed Nodes2STRNet. In Section 3, the proposed method is validated through shaking table tests on a four-story-building model. Conclusions are summarized in Section 4.

2 | METHODOLOGY

This study accomplishes dense structural displacement recognition based on the following ideas. First, the motion of structural control nodes in DSMM represents the motion of structure; therefore, a NodesEstimate subnetwork is established to generate 2D position coordinates of control nodes from the video frame as the input based on dense optical flow. Then, a Nodes2PoseNet subnetwork is established to model the mapping relationship between control node coordinates and structural pose parameters. Finally, the structural 3D mesh model is deformed according to the estimated structural pose parameters, and the dense displacement of the structure is obtained.

An overall schematic of the proposed Nodes2STRNet is shown in Figure 1. The workflow can be completed through three steps as follows:

Step 1: The NodesEstimate subnetwork uses each frame in the original video of a structure as input, outputs the 2D dense optical flow field compared with the original frame, and calculates the 2D control node heatmap of each frame.

Step 2: The 2D control node coordinates are converted from the control node heatmap of each frame by calculating the centroid and inputted into the Nodes2PoseNet subnetwork to obtain the structural pose parameters.

Step 3: The structural pose parameters are utilized to generate the DSMM in each frame, and the dense displacement of the structure is finally refined from the coordinates of vertices in DSMM.

Section 2 is organized following a logical order as described below. Before the methodology details, explanations of how the

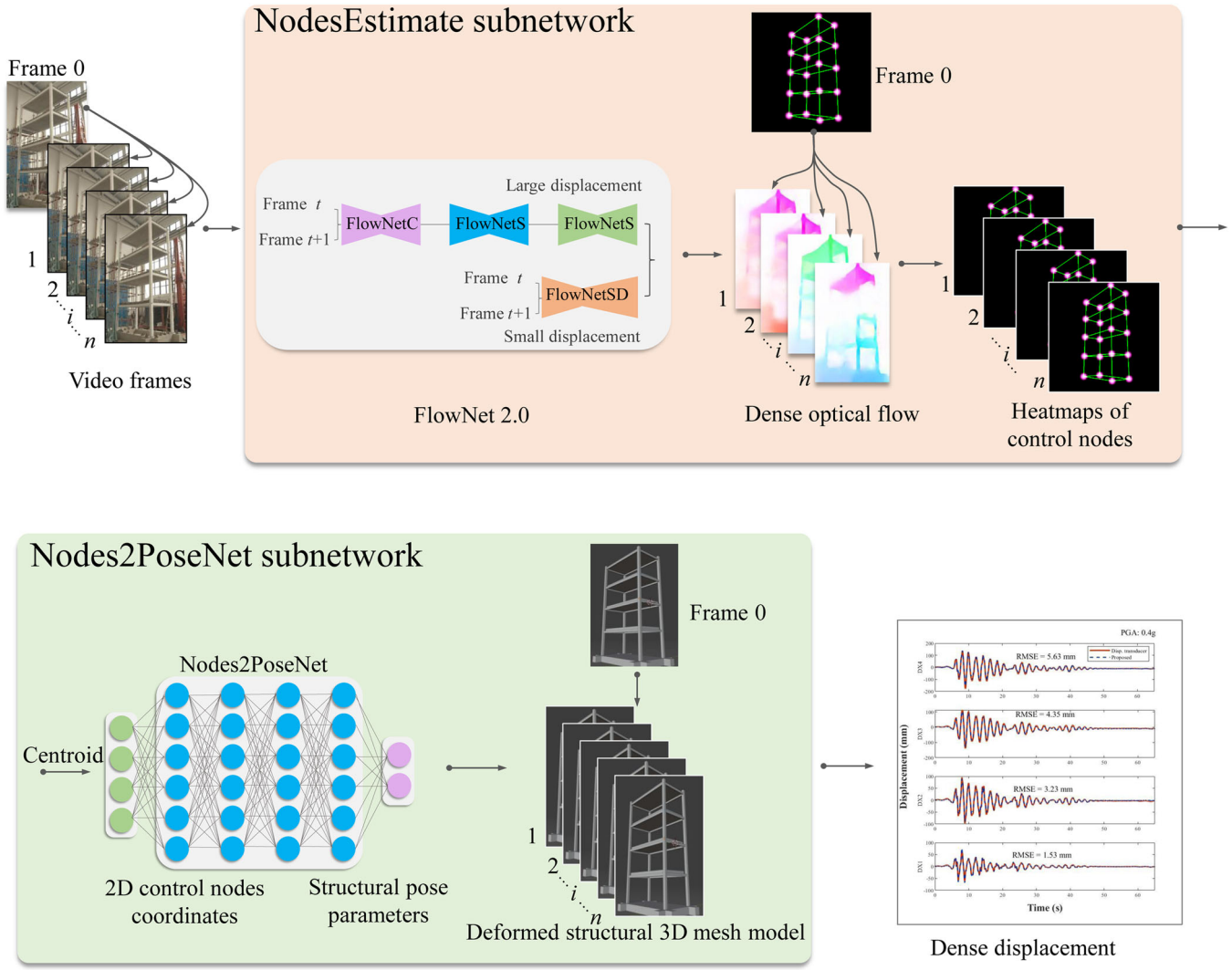

FIGURE 1 Identification workflow of proposed Nodes2STRNet. NodesEstimate, control node estimation subnetwork; Nodes2PoseNet, pose parameter recognition subnetwork.

method is proposed and the overall framework are introduced. Then, Section 2.1 shows how a DSMM is established for a structure as it forms the geometry foundation of the proposed method, and Section 2.2 introduces several key concepts and variable definitions for the input and output quantification, mainly including control nodes, 2D heatmaps, and coordinate transformation. Afterward, Sections 2.3 and 2.4 describe the network structure of the NodesEstimate and Nodes2PoseNet, respectively. Finally, Section 2.5 illustrates the efficient self-supervised training strategy of the proposed method.

2.1 | Deformable structural 3D mesh model

DSMM categorizes a structure into 3D elements, and each element is represented by a 3D deformable mesh model. The mesh model in an oscillation sequence is divided into the initial mesh model M_0 (as shown in Figure 2) and the time-variant mesh model M_t with pose parameter P_t (defined later). The basic elements of the initial mesh model M_0 include vertices V_0. A set of adjacent vertices is connected with each other by edges E to form a face. All faces form the surface of the mesh model M_0 and can be expressed as an undirected graph G with vertices V_0 and edges E:

$$M_0 = G(V_0, E). \tag{1}$$

At the frame t, M_t shares the same undirected graph G with M_0. The coordinates of vertices V_t shift with the oscillation of the structure. Therefore, M_t is an undirected graph consisting of new vertices V_t and the same edge E:

$$M_t = G(V_t, E). \tag{2}$$

A set of vertices $V_{S_i,t} \in V_t$ inside each cross-section of the structural component forms a common section $S_{i,t}$, $i = 1, 2, \ldots, n_a$, which can also be regarded as an undirected graph:

$$S_{i,t} = G_{S_i}(V_{S_i,t}, E_{S_i}), \quad i = 1, 2, \ldots, n_a, \tag{3}$$

where n_a is the number of common sections and determined by the vertex interval in the perpendicular direction of cross-sections, G_{S_i} is a subgraph of G, thus $S_{i,t} \subset M_t$.

Structural pose parameters $P_{S_{i,t}}$ control the motion of $S_{i,t}$, as shown in Figure 3. A few specific sections S_{C_i}, $i = 1, 2, \ldots, n_b$ with equal intervals are selected as the control section, where n_b is a hyperparameter for the number of control sections. The selection of n_b is a trade-off between the prediction accuracy of dense displacement and the model parameter volume of Node2STRNet.

Structural pose parameters $P_{S_{i,t}}$ consist of the transition $H_{S_{i,t}}$ and rotation angle $R_{S_{i,t}}$:

$$S_{i,t} = T(H_{S_{i,t}}, R_{S_{i,t}}; S_{i,0}), \tag{4}$$

where $S_{i,0}$ is the initial state of the common section $S_{i,t}$, and T is the conversion function with transition and rotation transformations, assuming that the common section $S_{i,t}$ is rigid. Note that $P_{S_{i,t}}$ in common sections S_i are calculated by cubic spline interpolation of $P_{S_{C_i,t}}$ in control sections S_{C_i}. ($P_{S_{C_i,t}}$ is predicted by Nodes2 PoseNet introduced in Section 2.4 later.) Combining Equations (1)–(4), M_t can be determined after all common sections S_i are obtained:

$$S_{i,t} = T[H_{S_{i,t}}, R_{S_{i,t}}; G_{S_i}(V_{S_{i,0}}, E_{S_i})] \subset M_t. \tag{5}$$

Considering that G_{S_i} and E_{S_i} are time-invariant and $V_{S_{i,0}}$ is known, M_t is determined by $H_{S_{i,t}}$ and $R_{S_{i,t}}$. In summary, Figure 3 shows the schematics of structural pose parameters in DSMM.

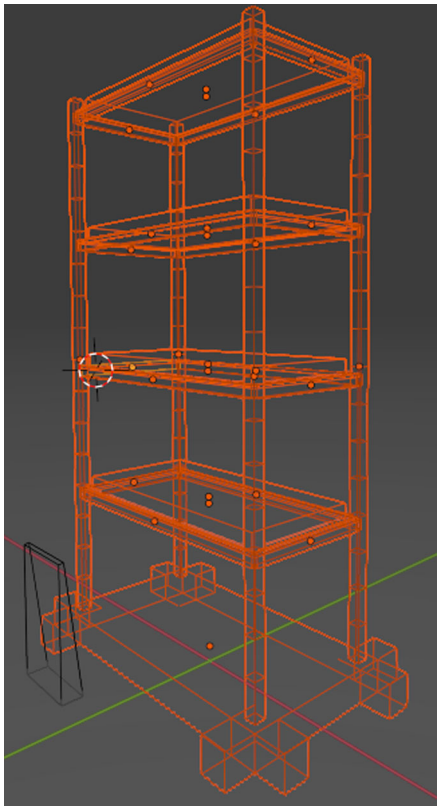

FIGURE 2 Initial structural 3D model of a four-story building.

2.2 | Definition of control nodes and 2D heatmaps

In Sections 2.2 and 2.3, the NodesEstimate subnetwork is designed to calculate the 2D control node heatmaps from each video frame. In Section 2.4, control node heatmaps are fed into the proposed Nodes2PoseNet subnetwork to obtain the structural pose parameters. Therefore, structural control nodes are used as intermediate connections between two subnetworks of Nodes2STRNet.

Heatmaps of structural control nodes represent structural motion and can be calculated by dense optical flow. Nodes2PoseNet uses FlowNet 2.0[32] to extract dense optical flow, representing the velocity field in a video frame. Compared to semantic segmentation masks, the dense optical flow contains sufficient information density because of the pixel-level velocity gradient.

Control nodes $N_{c3}(X, Y, Z)$ have a clear physical meaning of spatial location in real-world 3D coordinates, as shown in Figure 4. Control nodes are usually set on the joints of structural elements, for example, joints of columns and beams at each story of a frame structure. A higher density of control nodes can result in higher

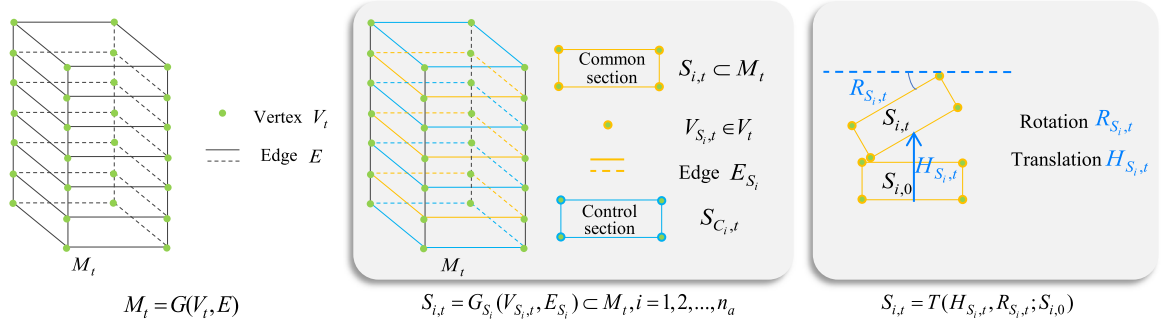

FIGURE 3 Schematics of structural pose parameters in DSMM with common and control sections. DSMM, deformable structural 3D mesh model.

spatial resolution of the 3D mesh model, which improves the identification accuracy of dense structural displacement.

To calculate the 2D coordinates of control nodes $N_{c2}(x, y)$ in the initial frame, the 3D coordinates of control nodes in the initial 3D mesh model and the camera matrix are required:

$$N_{c2} = R_c N_{c3} + T_c, \tag{6}$$

where R_c and T_c represent the rotation matrix and translation matrix of the camera, respectively.

The 2D heatmaps in the initial frame Hm_0^i are generated around 2D control nodes following a normalized 2D Gaussian distribution:

$$Hm_0^i = \begin{cases} N\left(N_{c2}^i, \sigma\right), & N\left(N_{c2}^i, \sigma\right) \geq n_{\text{th}}, \\ 0, & N\left(N_{c2}^1, \sigma\right) < n_{\text{th}}, \end{cases} \tag{7}$$

where i is the index of control nodes $i = 1, 2..., N_{\text{nodes}}$, N_{nodes} is the number of control nodes, $N(N_{c2}^i, \sigma)$ is the probability density function

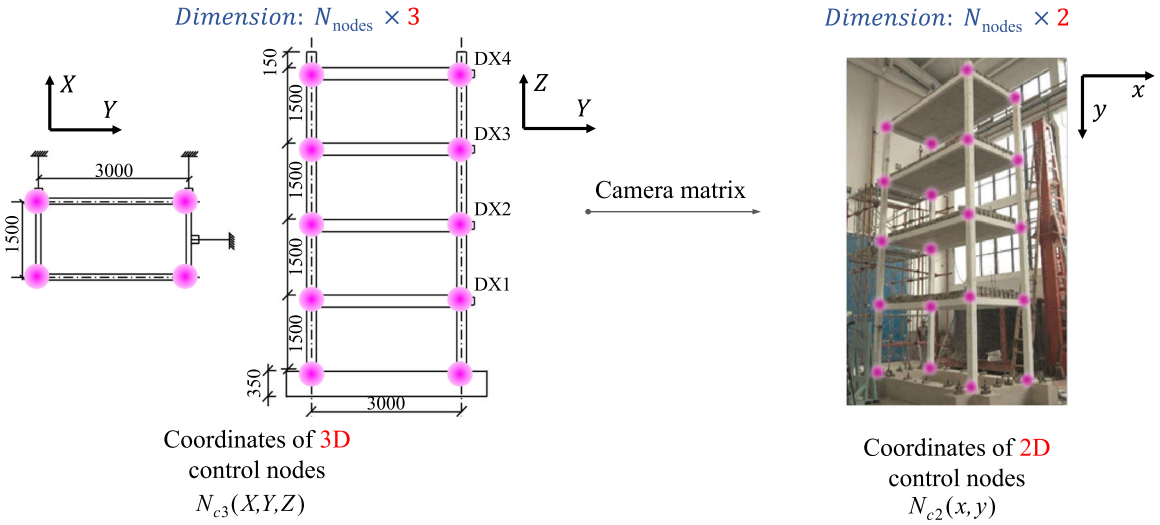

FIGURE 4 Control node transformation from 3D real-world coordinates to 2D image coordinates through the camera matrix.

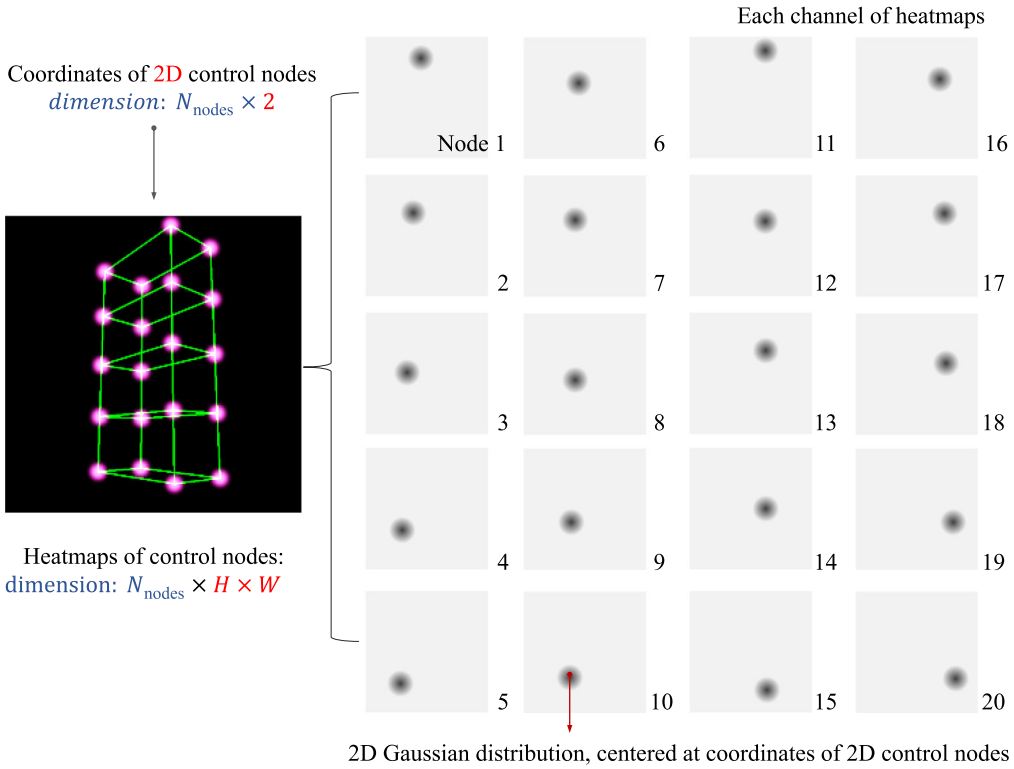

FIGURE 5 Heatmaps of control nodes in the initial video frame.

of normalized 2D Gaussian distribution, N_{c2}^{i} denotes the 2D coordinates of the ith control nodes in the initial video frame, σ is the standard deviation representing the pixel range around control nodes, and n_{th} is the preset threshold value. The schematics of generating the heatmaps from 2D control nodes are shown in Figure 5.

2.3 | NodesEstimate subnetwork

The NodesEstimate subnetwork utilizes each video frame as input, predicts the dense optical flow between each frame and the initial frame by FlowNet 2.0, and then calculates the heatmaps of 2D control nodes for each frame, as shown in Figure 1. The dense optical flow represents the pixel-level displacement vector field between the original frame and the subsequent image. As shown in the left part of Figure 1, the optical flow of two adjacent frames can be obtained by FlowNet 2.0. The optical flow field can be converted into RGB visualization through the color wheel conversion, where the modulus and direction of the optical flow vector are denoted as saturation and hue, respectively, as shown in Figure 6. By applying the predicted dense optical flow field to the heatmaps of control nodes in the initial frame, the corresponding heatmaps of control nodes in the following frames can be obtained, as shown in the right part of Figure 1.

The heatmap of ith control node in the tth frame Hm_t^i is obtained from

$$Hm_t^i = \text{NodesEstimate}\left[\text{Flow}\left(I_t, I_0\right), Hm_0^i\right], \tag{8}$$

where I_t, I_0 denote the tth and initial frames, Flow denotes the optical flow calculation process (FlowNet 2.0[32] in this study), Hm_0^i denotes the heatmap of the ith control node in the initial frame, and can be obtained from Equation (7) in Section 2.2. The NodesEstimate subnetwork only performs the feedforward interference process to calculate the control node heatmaps using the pretrained FlowNet 2.0.

2.4 | Nodes2PoseNet subnetwork

After the 2D coordinates are obtained by the NodesEstimate subnetwork, the Nodes2PoseNet subnetwork is established to generate structural pose parameters from the 2D coordinates of control nodes.

The heatmap centroid coordinate of the ith control node in the tth frame can be calculated using the equations:

$$N_{c2,t} = \left[CH_t^i, CW_t^i\right],$$
$$CH_t^i = \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} x \times Hm_t^i(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W} Hm_t^i(x,y)}, \quad CW_t^i = \frac{\sum_{x=1}^{H}\sum_{y=1}^{W} y \times Hm_t^i(x,y)}{\sum_{x=1}^{H}\sum_{y=1}^{W} Hm_t^i(x,y)}, \tag{9}$$

where H and W, respectively, denote the height and width of the frame and are indexed by x and y, Hm_t^i denotes the heatmap of the ith control node in the tth frame and can be obtained using Equation (8) in Section 2.3, CH_t^i and CW_t^i denote the centroid coordinates $N_{c2,t}^i$ of ith control node in the tth frame in the height and width directions, respectively.

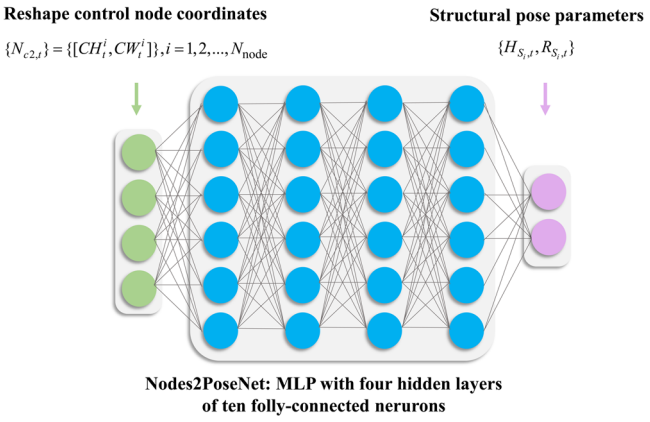

FIGURE 7 MLP architecture of Nodes2PoseNet subnetwork. MLP, multilayer perceptron.

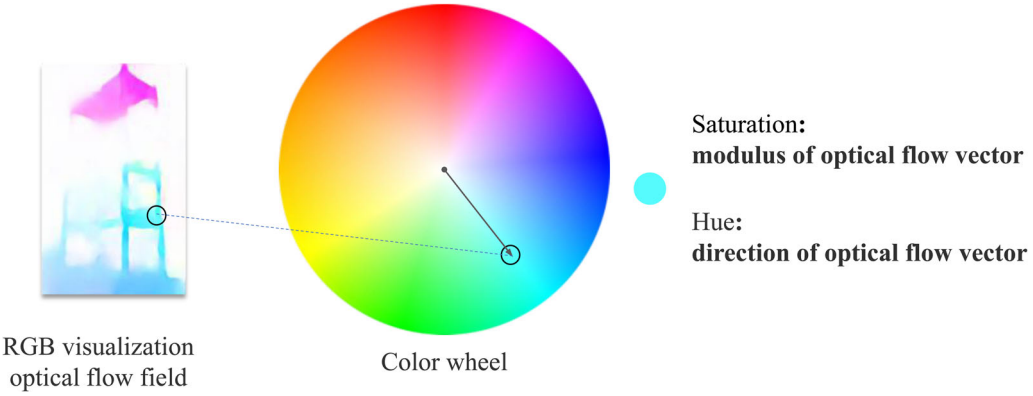

FIGURE 6 Color wheel conversion for RGB visualization of optical flow with arrow and modulus annotation.

As shown in Figure 1, the Nodes2PoseNet subnetwork utilizes the heatmap centroid coordinates as input and predicts the structural pose parameters as output:

$$\left\{H_{S_i,t}, R_{S_i,t}\right\} = \text{Nodes2PoseNet}\left\{N^i_{c2,t}\right\} \tag{10}$$

The Nodes2PoseNet subnetwork utilizes a network architecture of multilayer perceptron (MLP) with four hidden layers (as shown in Figure 7) and a self-supervised training strategy without manual annotations (details about the self-supervised training strategy will be explained in Section 2.5). According to Equation (10), each video frame can generate a pair of input (2D centroid coordinates of control nodes

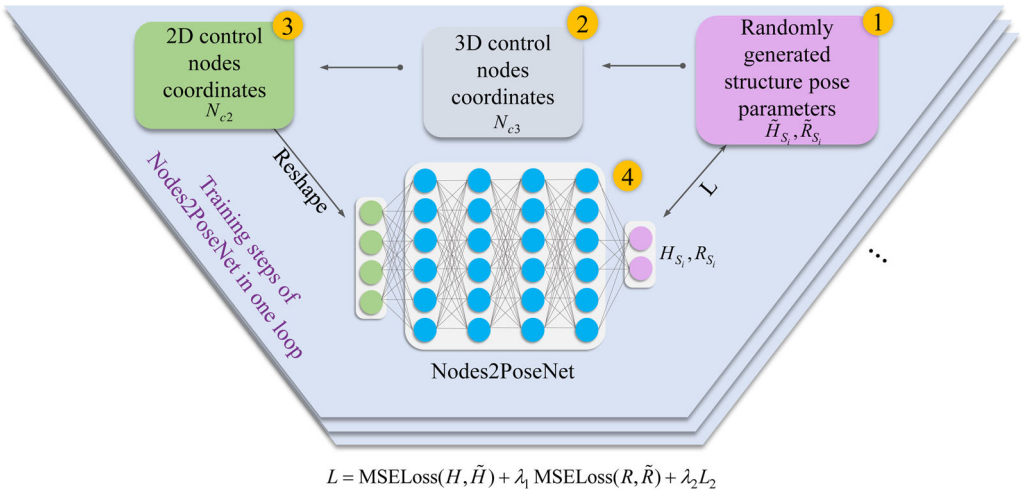

FIGURE 8 Overall schematic of a self-supervised training strategy for Nodes2PoseNet. Nodes2PoseNet, pose parameter recognition subnetwork.

TABLE 1 Earthquake ground motion with various intensities of shaking table tests.

	BM16	BM18	BM19	BM22	BM25
Earthquake ground motion	Shanghai artificial wave	El Centro	Wenchuan	Artificial wave	Wenchuan
Intensity	0.2*g*	0.4*g*	0.4*g*	0.6*g*	0.6*g*

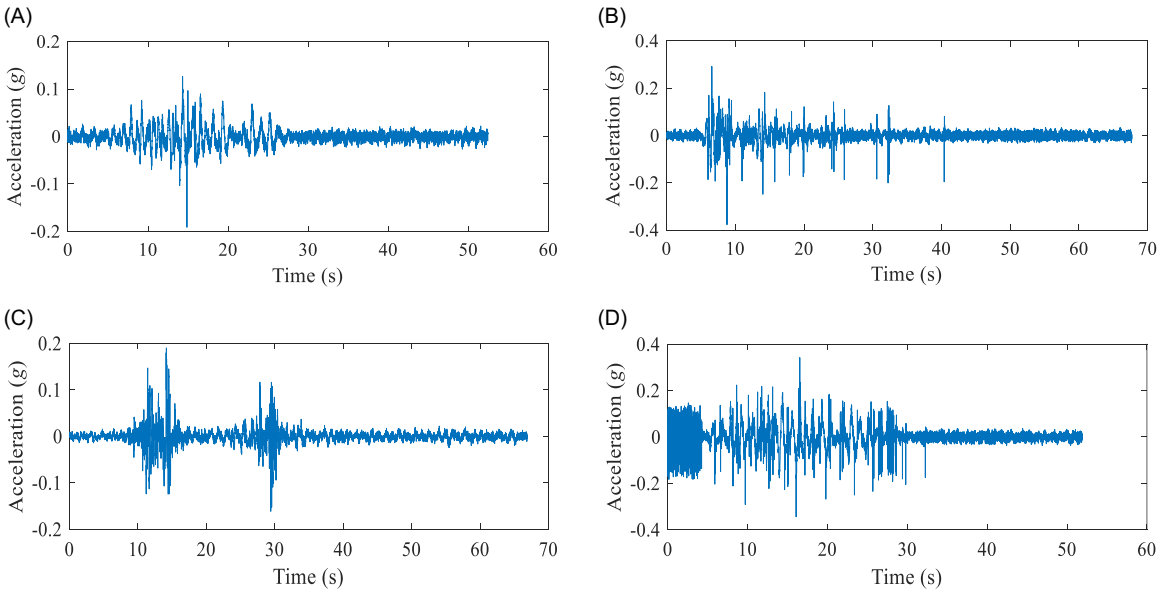

FIGURE 9 Waveforms of selected earthquake ground motions in shaking table tests. (A) BM16 Shanghai artificial wave, (B) BM18 El Centro, (C) BM19/BM25 Wenchuan, and (D) BM22 artificial wave.

TABLE 2 The 3D real-world and 2D image coordinates of control nodes in the initial structural model.

Control node No.	Story	3D coordinates (unit: meter)			2D coordinates (unit: pixel)	
		X	*Y*	*Z*	*x*	*y*
1	Bottom	1.5	0.35	0.75	292.176	462.312
2	Story 1	1.5	1.85	0.75	290.5061	335.5144
3	Story 2	1.5	3.35	0.75	288.9835	219.9069
4	Story 3	1.5	4.85	0.75	287.5896	114.0707
5	Story 4	1.5	6.35	0.75	286.3087	16.81725
6	Bottom	−1.5	0.35	0.75	160.3105	441.8237
7	Story 1	−1.5	1.85	0.75	163.6637	346.2861
8	Story 2	−1.5	3.35	0.75	166.7899	257.2177
9	Story 3	−1.5	4.85	0.75	169.7114	173.9829
10	Story 4	−1.5	6.35	0.75	172.4476	96.02681
11	Bottom	−1.5	0.35	−0.75	234.4249	435.8794
12	Story 1	−1.5	1.85	−0.75	235.1104	349.4524
13	Story 2	−1.5	3.35	−0.75	235.7536	268.348
14	Story 3	−1.5	4.85	−0.75	236.3584	192.0891
15	Story 4	−1.5	6.35	−0.75	236.9282	120.2543
16	Bottom	1.5	0.35	−0.75	371.4525	452.0849
17	Story 1	1.5	1.85	−0.75	366.7597	340.8523
18	Story 2	1.5	3.35	−0.75	362.4335	238.309
19	Story 3	1.5	4.85	−0.75	358.4325	143.4752
20	Story 4	1.5	6.35	−0.75	354.7214	55.51296

with a dimension of $1 \times 2N_{nodes}$) and output (with a dimension of structural pose parameters in all control sections). Ten neurons are equally included in each of the four hidden layers.

Compared with the Structure-PoseNet architecture in the previous study,[30] the Nodes2PoseNet subnetwork can be directly transferred from the training data sets to the actual recognition scenarios. Artificially generated data in ParaNet of Structure-PoseNet cannot perfectly simulate some real-world scenarios because of slight differences in structural morphology and prediction variations between semantic segmentation masks from synthetical environments and actual videos. Therefore, prediction errors exist in real-world recognition using Structure-PoseNet from the training data. As a comparison, the Nodes2PoseNet subnetwork utilizes the centroid coordinates of a few control nodes as input, which only includes

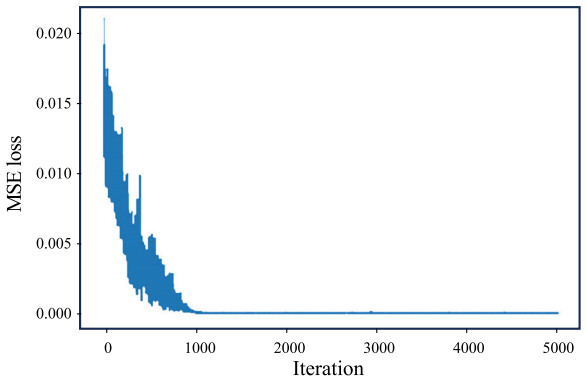

FIGURE 11 Training loss descending curve for Nodes2PoseNet subnetwork. MSE, mean-square error; Nodes2PoseNet, pose parameter recognition subnetwork.

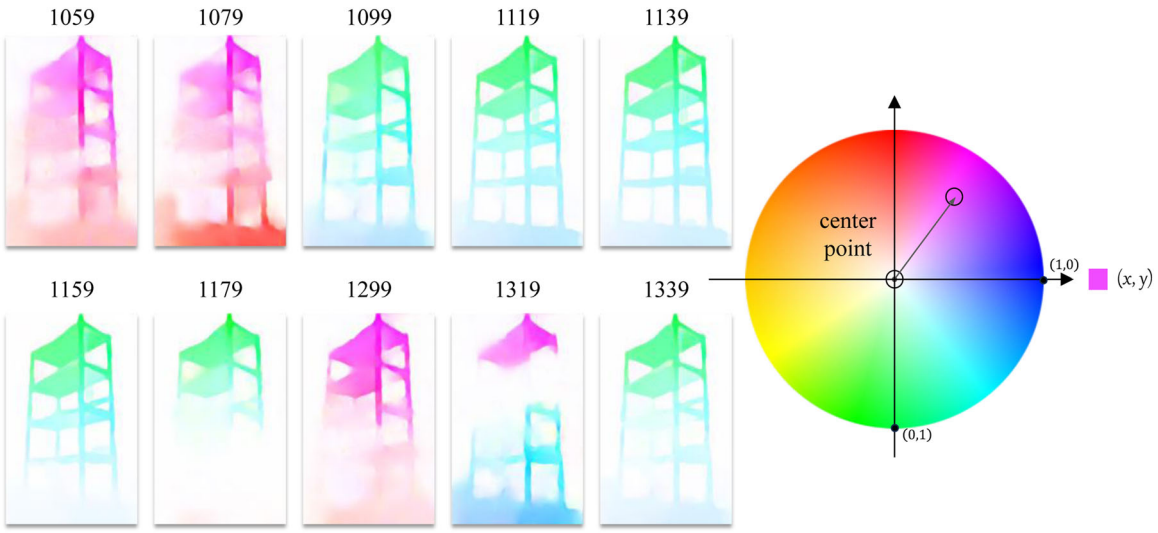

FIGURE 10 Representative recognition results of dense optical flow by NodesEstimate subnetwork. NodesEstimate, control node estimation subnetwork.

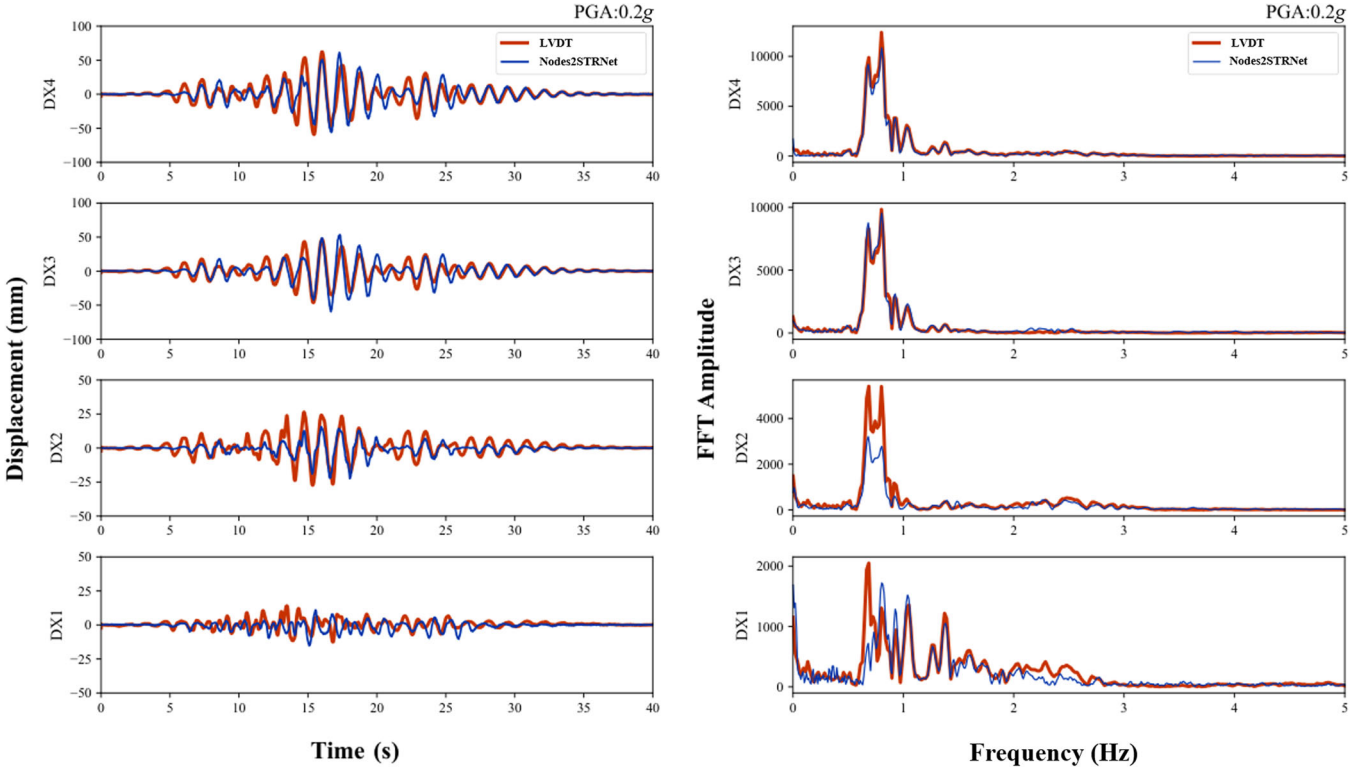

FIGURE 12 Comparison results of recognized and LVDT displacement time-histories and frequency distributions in BM16. FFT, fast Fourier transform; LVDT, linear variable differential transformer; PGAs, peak ground accelerations.

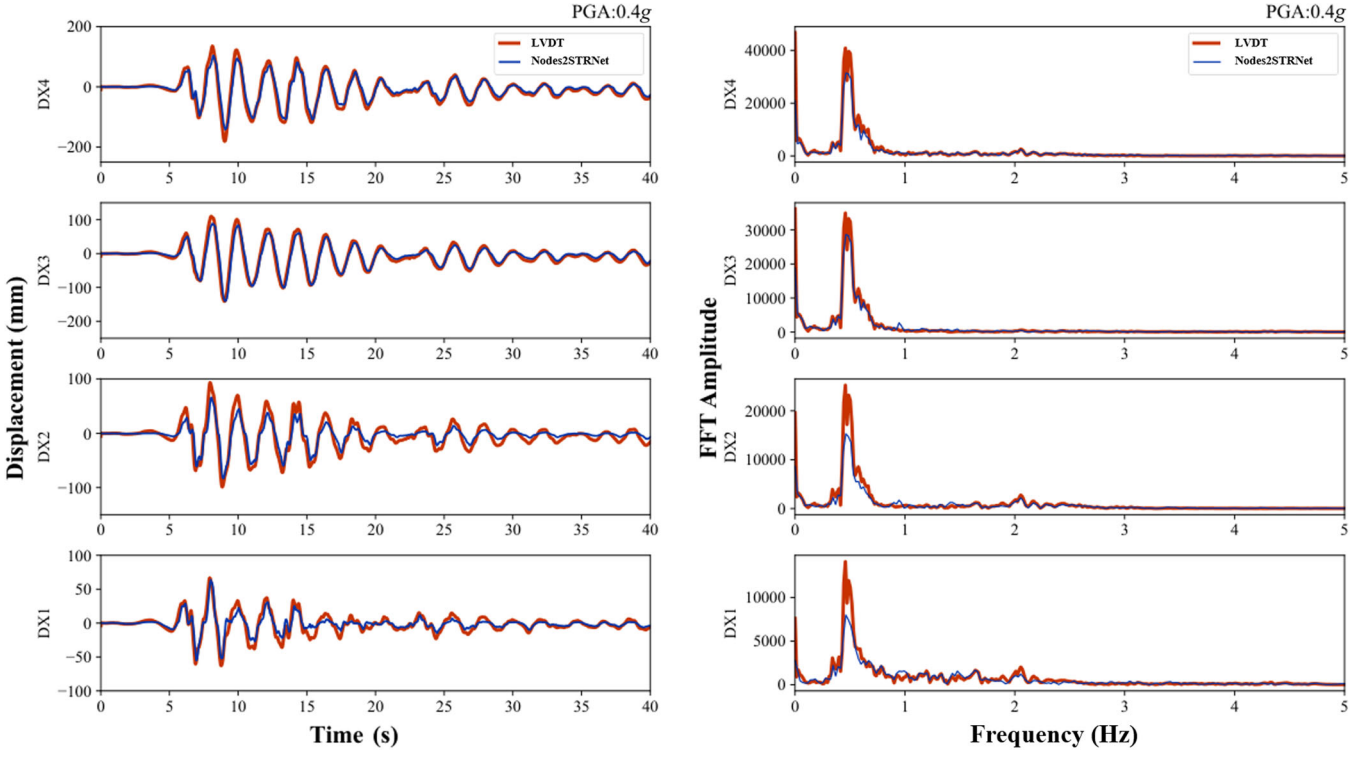

FIGURE 13 Comparison results of recognized and LVDT displacement time-histories and frequency distributions in BM18. FFT, fast Fourier transform; LVDT, linear variable differential transformer; PGAs, peak ground accelerations.

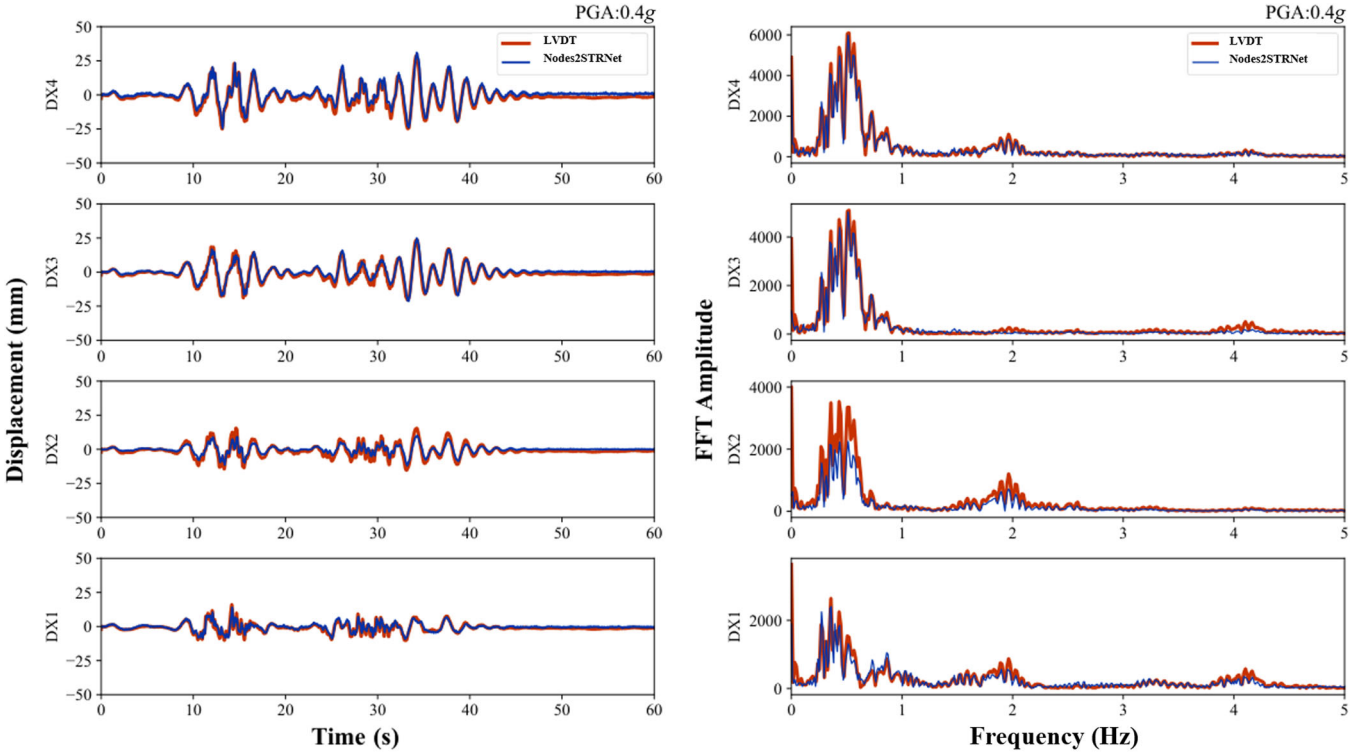

FIGURE 14 Comparison results of recognized and LVDT displacement time-histories and frequency distributions in BM19. FFT, fast Fourier transform; LVDT, linear variable differential transformer; PGAs, peak ground accelerations.

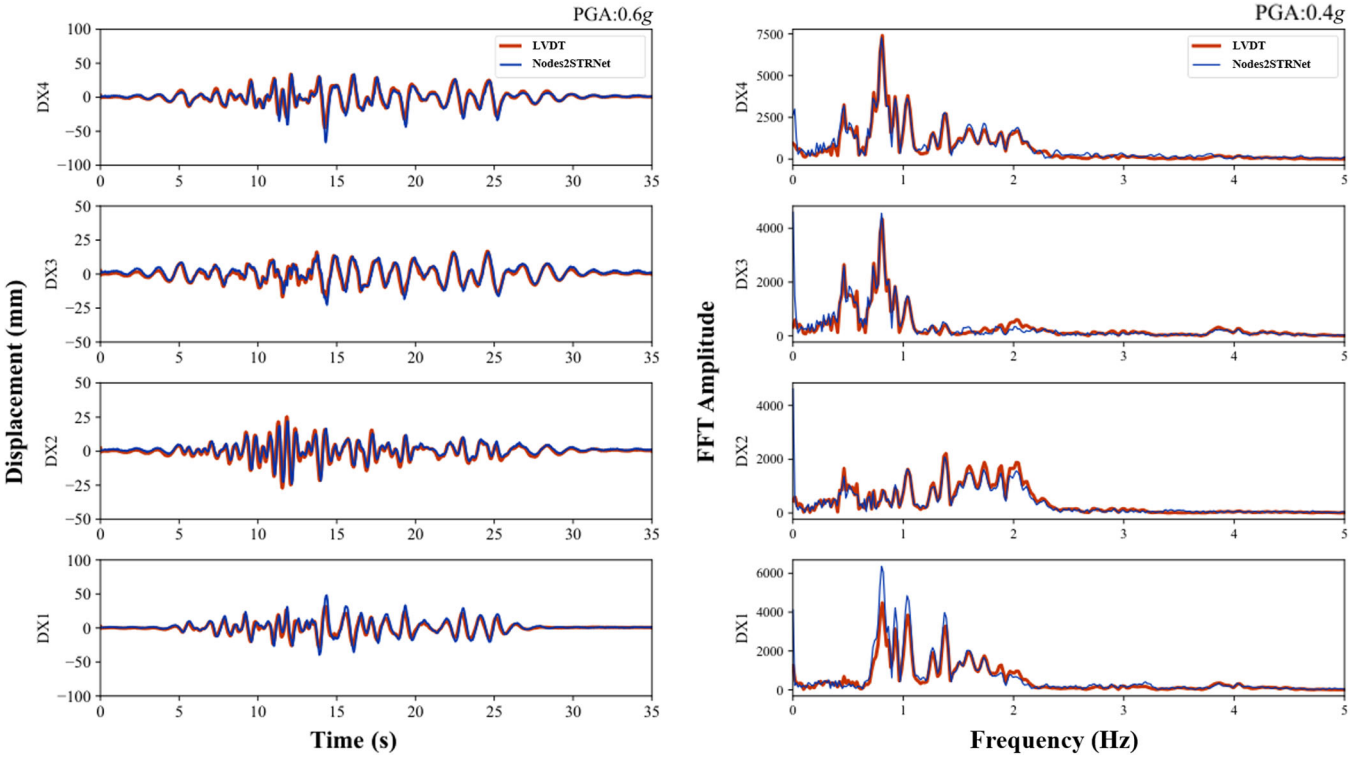

FIGURE 15 Comparison results of recognized and LVDT displacement time-histories and frequency distributions in BM22. FFT, fast Fourier transform; LVDT, linear variable differential transformer; PGAs, peak ground accelerations.

necessary information and avoids uncontrollable noise compared to semantic segmentation masks of all pixels.

2.5 | Self-supervised training strategy of Nodes2Posenet

Figure 8 shows the overall schematic of the self-supervised training process for the Nodes2PoseNet subnetwork, including the following steps:

(i) Structural pose parameters $\hat{H}$, $\hat{R}$ are randomly generated following a uniform distribution in a fixed range as the ground-truth values:

$$\hat{H} = \text{Uniform}(H_a, H_b), \quad \hat{R} = \text{Uniform}(R_a, R_b), \tag{11}$$

where H_a, H_b, R_a, R_b denote the preset lower and upper bounds for structural pose parameters H, R, and "Uniform" denotes the uniform distribution function.

(ii) The randomly generated structural pose parameters $\hat{H}$, $\hat{R}$ are applied to the initial structural 3D model, and the 3D spatial coordinates of all control nodes can be obtained according to Equation (5).

(iii) The 3D spatial coordinates are converted into 2D camera coordinates by camera matrix transformation to generate the 2D coordinates of control nodes according to Equation (6).

(iv) The flattened 2D coordinates of control nodes are utilized as input to the Nodes2PoseNet subnetwork according to Equation (10), output the predicted H and R, calculate the regression loss with the ground-truth values of $\hat{H}$ and $\hat{R}$, and update the Nodes2PoseNet subnetwork by the Adam algorithm.

The mean-square error (MSE) loss function with L_2 regularization is adopted to calculate the regression error between ground-truth structural pose parameters and predicted values corresponding to all the control nodes:

$$\begin{aligned} L &= \text{MSELoss}(H, \hat{H}) + \lambda_1 \text{MSELoss}(R, \hat{R}) + \lambda_2 L_2 \\ &= \frac{1}{N_{\text{nodes}}} \sum_{i=1}^{N_{\text{nodes}}} \left[(H_i - \hat{H}_i)^2 + \lambda_1 (R_i - \hat{R}_i)^2 \right] + \lambda_2 L_2, \end{aligned} \tag{12}$$

where L denotes the loss function to update the Nodes2PoseNet subnetwork, N_{nodes} denotes the number of control nodes, λ_1 denotes the weight coefficient between H and R, L_2 denotes the regularization term to avoid overfitting, and λ_2 denotes the regularization coefficient.

(v) Return to the first step, repeat steps (i)–(iv), and iteratively update Nodes2PoseNet until the loss reduces below a preset value ε (set as 10^{-5} in this study).

3 | VALIDATION EXPERIMENT

3.1 | Experimental setups of shaking table test and network training

To verify the identification accuracy of dense displacement and the robustness to different light conditions, a seismic shaking table test of a four-story-building model is conducted. As shown in Figure 4, four

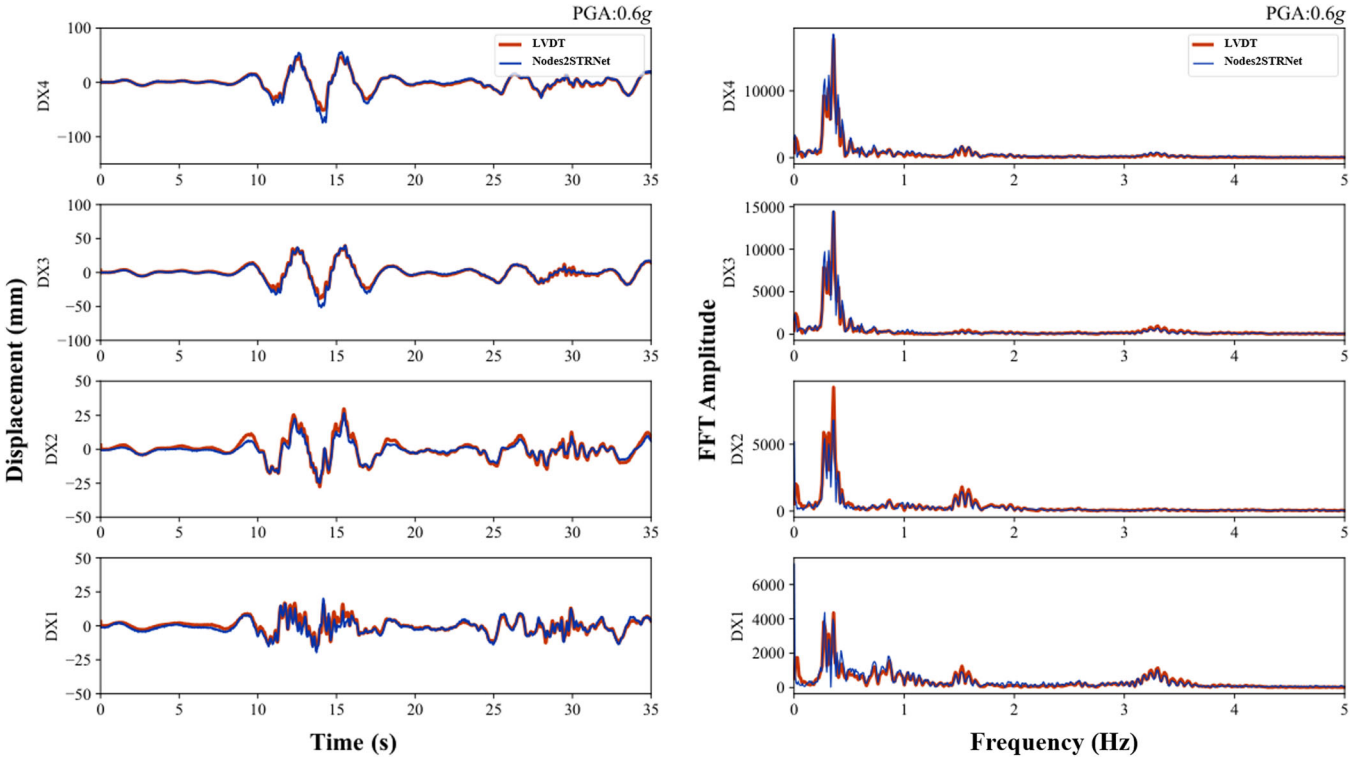

FIGURE 16 Comparison results of recognized and LVDT displacement time-histories and frequency distributions in BM25. FFT, fast Fourier transform; LVDT, linear variable differential transformer; PGAs, peak ground accelerations.

LVDTs (i.e., DX1, DX2, DX3, and DX4) were installed to measure the displacement in the vibration direction with a sampling frequency of 256 Hz. A fixed camera was set at a distance of 8 m from the building model with a video frame rate of 60 Hz and an original resolution of 2160 × 3840. Five videos, namely, BM16, BM18, BM19, BM22, and BM25, were recorded under different earthquake ground motions and light illuminations. Table 1 shows the investigated earthquake ground motions with their intensities of the five videos during the shaking table tests, and the corresponding waveforms are shown in Figure 9. More details can be found in previous studies.[22,30]

As shown in Figure 5, a total of 20 control nodes are set on the junction points of the ground and the first floor and joints of columns and beams at each story. The 3D coordinates of these control points are converted into 2D image coordinates by the known camera matrix, as shown in Table 2.

Each video frame is an input to the NodesEstimate subnetwork, which predicts the dense optical flow between itself and the initial frame, and this predicted dense optical flow field is converted into RGB images for visualization through color wheel conversion. Some representative recognition results of dense optical flow by the NodesEstimate subnetwork are shown in Figure 10. The number above each subfigure represents the frame number in the video, and the structure motion direction can be intuitively observed by different hues in the RGB map of the optical flow field. The length of the optical flow vector is normalized to the range between 0 and 1.

The hyperparameters for training the Nodes2StrNet subnetwork are set as follows. The upper and lower bounds of the uniform distribution to randomly generated structural pose parameters H and R are $H_a = -0.2$, $H_b = 0.2$, and $R_a = R_b = 0$ due to the 1D motion direction of the shaking table. The Adam optimization is utilized with an initial learning rate of 0.0001, a batch size of 5, a total number of iterations inside an epoch of 100, and a training epoch of 50. A learning rate decay strategy is adopted, reducing by half in every 10 epochs. The loss descending curve during the training process of Nodes2PoseNet subnetwork is shown in Figure 11.

3.2 | Results and discussion

Figures 12–16 show the comparison results between the recognized displacement time-histories by the proposed Nodes2STRNet and measured displacement time-histories by LVDT at DX1–4 and the corresponding frequency transformation by the fast Fourier transform (FFT) algorithm. The results show that the recognized multistory displacements by the proposed Nodes2STRNet match well with those of LVDTs under different peak ground accelerations. The prediction error tends to be larger at the bottom of the structure, and one possible reason is that some parts at the bottom are missing in specific video frames, leading to the recognition inconsistency in actual videos using a well-trained model.

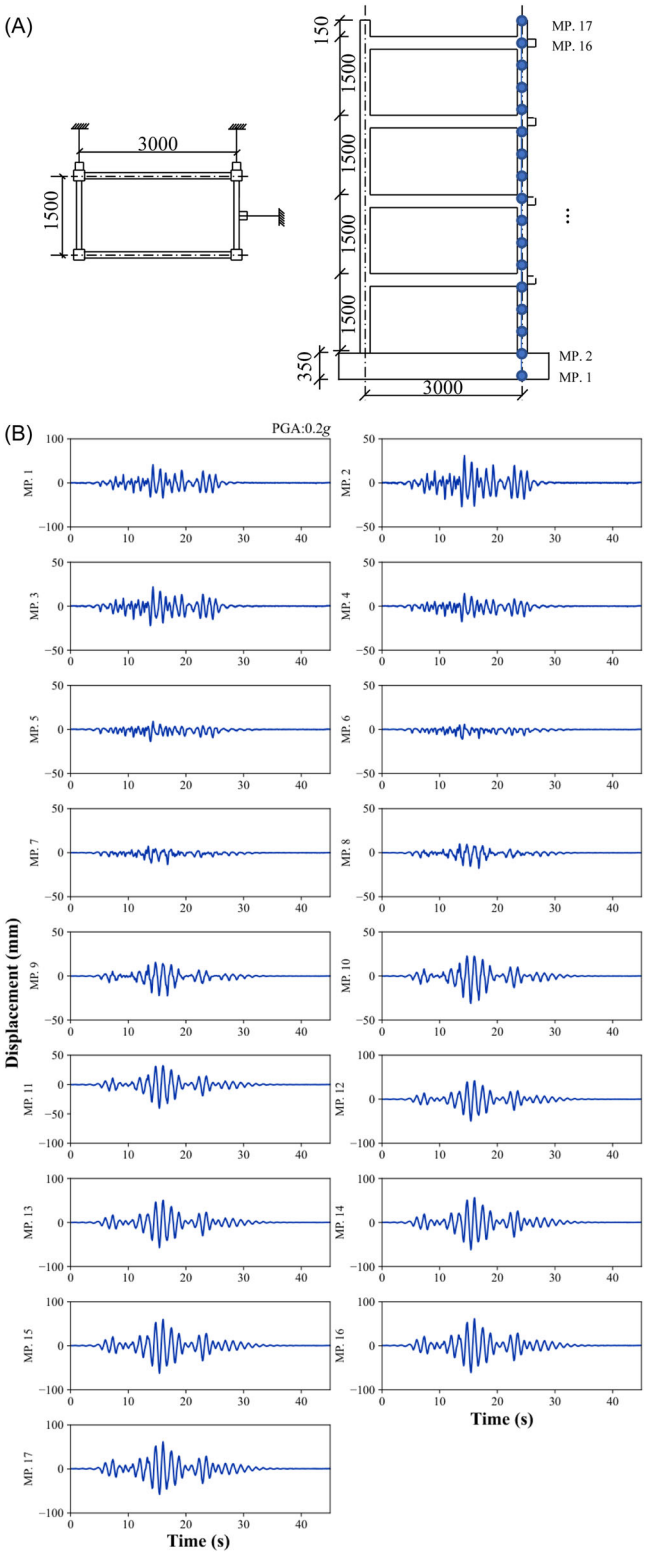

FIGURE 17 Dense displacement recognition results by proposed Nodes2STRNet in BM16. (A) Schematic of dense displacement measurement points of MP.1–17 (unit: millimeter). (B) Recognized time histories of dense displacement. PGA, peak ground acceleration.

Figure 17 shows the dense displacement recognition results by the proposed Nodes2STRNet in BM16, in which MP.1–17 represents the dense measurement points along the height direction of the structure. More results of BM18, BM19, BM22, and BM25 are shown in Figures A1–A4 of the appendix. The results verify that the proposed Nodes2STRNet can obtain dense structural dynamic displacement from a monocular vibration video.

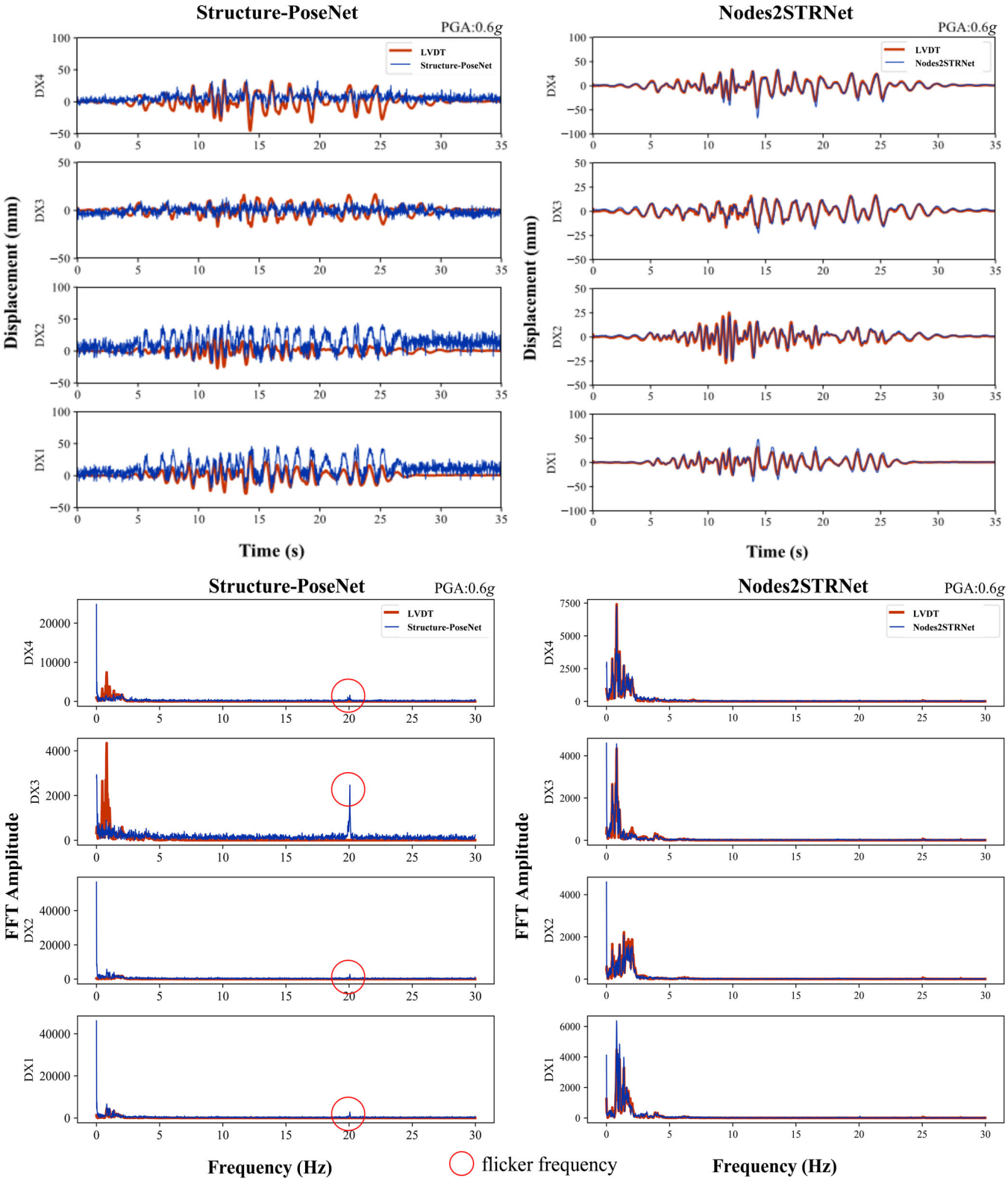

FIGURE 18 Comparisons of recognized multistory displacement histories and FFT results between Structure-PoseNet and proposed Nodes2STRNet (BM22). FFT, fast Fourier transform; LVDT, linear variable differential transformer; PGA, peak ground acceleration.

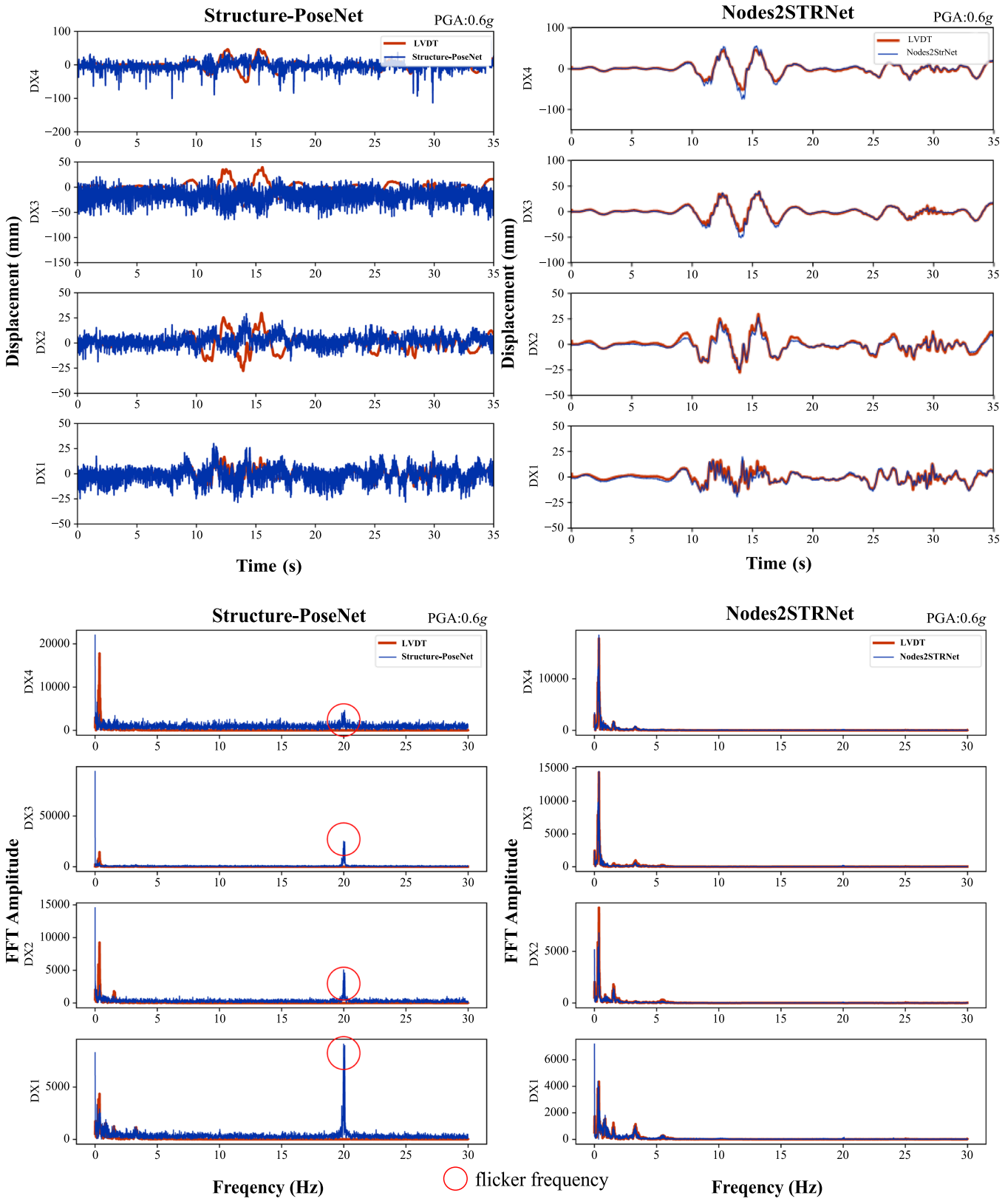

FIGURE 19 Comparisons of recognized multistory displacement histories and FFT results between Structure-PoseNet and proposed Nodes2STRNet (BM25). FFT, fast Fourier transform; LVDT, linear variable differential transformer; PGA, peak ground acceleration.

Figures 18 and 19 compare the recognized multistory displacement histories and the corresponding FFT results between Structure-PoseNet[30] and the proposed Nodes2STRNet in BM22 and BM25. The recognition results of Structural-PoseNet show noticeable prediction errors, while those from Nodes2STRNet match well with the measurements from LVDT. In addition, the FFT results of Structure-PoseNet have a noise peak at 20 Hz because of the video flicker effect,[33] while Nodes2StrNet has robustness against the variation of light conditions.

Figure 20 compares the effects of light condition variations on the model robustness of Structure-PoseNet and Nodes2STRNet. Apparent pixel-level noises exist in the semantic segmentation masks of Structure-PoseNet for BM25 compared to BM16. However, the corresponding control nodes are well identified by Nodes2STRNet with good robustness. The results validate that the proposed Nodes2StrNet is more robust to light condition variations among different video frames. Similarly, low video quality can cause significant recognition noise in semantic segmentation masks for Structure-PoseNet, affecting the recognition accuracy of dense structural displacements. As a comparison, the proposed Nodes2STRNet is flexible to the video resolution because the NodesEstimate subnetwork does not require a training process; however, Structure-PoseNet requires a fixed input resolution and should be retrained when faced with a new video with a different resolution.

Table 3 compares the average root-mean-square error (RMSE) and pixel-wise root-mean-square error (RMSE-PX) for DX1–4 using Structure-PoseNet[30] and the proposed Nodes2STRNet. The results show that for BM16, BM19, BM22, and BM25, the proposed Nodes2StrNet shows significant improvements, further validating its robustness against light condition variations and superiority to Structure-PoseNet.[30] For BM18, the possible reason may be that control nodes at the bottom of the structure exceed the image boundary, leading to the inaccurate calculation of coordinates. The results also indicate that the view field of the camera should be adjusted to ensure that the maximum range of structural motion is within each video frame during the structural vibration process, which may be achieved by increasing the object distance and decreasing the focus distance. Specifically, the corresponding strategy of selecting optimal internal and external camera parameters should be determined according to the properties of the structural model and earthquake ground motion in the shaking table test.

TABLE 3 Average errors of recognized displacements for DX1–4 using Structure-PoseNet and Nodes2STRNet.

	BM16	BM18	BM19	BM22	BM25
Average RMSE of DX1–4					
Structure-PoseNet (mm)[30]	6.81	4.21	3.48	12.49	16.09
Nodes2STRNet (proposed) (mm)	6.79	10.71	2.21	4.31	2.46
Average RMSE-PX of DX1–4					
Structure-PoseNet (px)[30]	0.43	0.27	0.25	1.23	1.60
Nodes2STRNet (proposed) (px)	0.43	0.69	0.16	0.43	0.25

Abbreviations: RMSE, root-mean-square error; RMSE-PX, pixel-wise root-mean-square error.

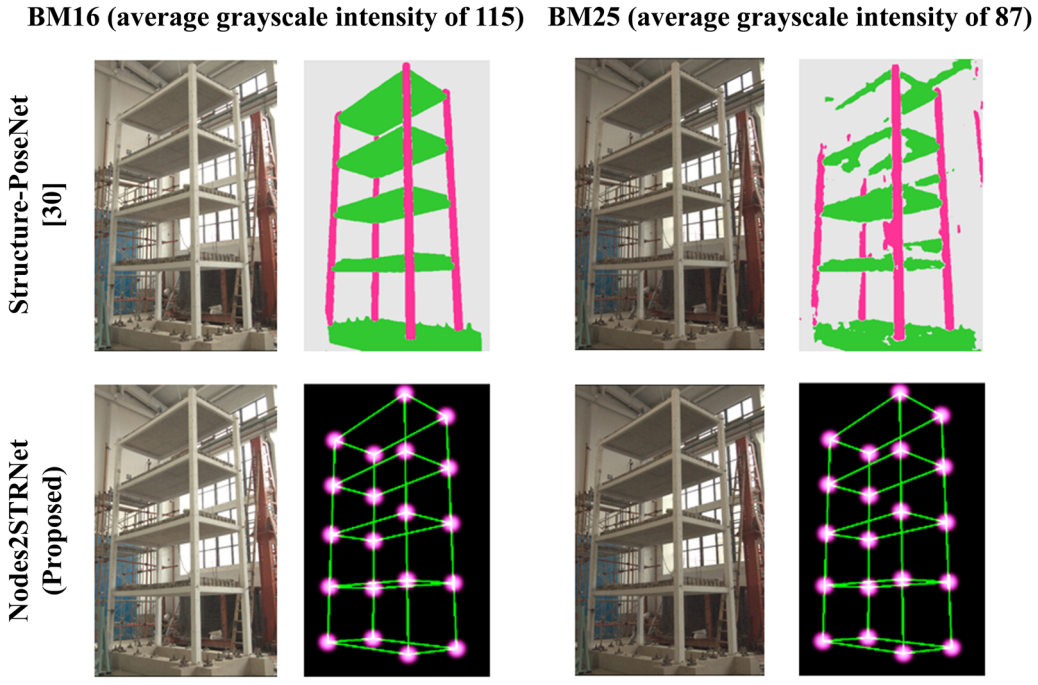

FIGURE 20 Effects of light condition variation on model robustness of Structure-PoseNet and Nodes2STRNet.

4 | CONCLUSION

This study proposes a novel Nodes2STRNet for structural dense displacement recognition based on DSMM and motion representation. The main conclusions are summarized as follows:

The proposed Nodes2STRNet comprises two subnetworks of NodesEstimate and Nodes2PoseNet. The NodesEstimate subnetwork utilizes each video frame as input, generates the dense optical flow field based on FlowNet 2.0, and outputs structural control nodes. The Nodes2PoseNet uses structural control nodes as input and predicts the structural pose parameters using an MLP.

Various DSMMs are generated according to structural pose parameters with motion representation for the entire structure, in which structural control nodes are utilized as intermediate connections between two subnetworks. The dense displacements of the structure can finally be obtained from DSMMs in different video frames.

A self-supervised learning strategy is designed to train the Nodes2PoseNet subnetwork. An MSE loss with L_2 regularization is adopted to calculate the regression error between the ground-truth and predicted structural pose parameters of all control nodes.

The recognition effectiveness and accuracy of dense displacement and robustness to light condition variations are validated by shaking table test of a four-story frame structure scale model. The results show that the average RMSE-PX of multistory displacement histories using the proposed Nodes2STRNet ranges from 0.16 to 0.69 pixels.

Compared with image-segmentation-based Structure-PoseNet using all pixel information, the proposed Nodes2STRNet gains higher robustness against light condition variations with a more straightforward self-supervised training process using only a few control nodes. In addition, the proposed Nodes2STRNet obtains higher flexibility in the input video resolution because the NodesEstimate subnetwork applies a pretrained FlowNet 2.0 to generate the dense optical flow field without additional training faced with new scenarios.

In addition, the effects of the *P* wave and *S* wave on the structural dense displacement recognition based on DSMM and motion representation by the proposed Nodes2STRNet will be further investigated in a future study.

ACKNOWLEDGMENTS

Financial support for this study was provided by the National Natural Science Foundations of China (Grant Nos. 52192661, 51921006, and 52008138), China Postdoctoral Science Foundations (Grant Nos. BX20190102 and 2019M661286), Heilongjiang Provincial Natural Science Foundation (Grant No. LH2022E070), and Heilongjiang Province Postdoctoral Science Foundations (Grant Nos. LBH-TZ2016 and LBH-Z19064).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Yang Xu http://orcid.org/0000-0002-8394-9224

REFERENCES

1. Olaszek P. Investigation of the dynamic characteristic of bridge structures using a computer vision method. *Measurement*. 1999;25(3):227-236.
2. Lee JJ, Shinozuka M. A vision-based system for remote sensing of bridge displacement. *NDT E Int*. 2006;39(5):425-431.
3. Chi S, Caldas CH. Automated object identification using optical video cameras on construction sites. *Comput-Aided Civ Infrastruct Eng*. 2011;26(5):368-380.
4. Feng D, Feng MQ. Vision-based multipoint displacement measurement for structural health monitoring. *Struct Control Health Monit*. 2016;23(5):876-890.
5. Yang Y, Dorn C, Mancini T, et al. Blind identification of full-field vibration modes from video measurements with phase-based video motion magnification. *Mech Syst Signal Process*. 2017;85:567-590.
6. Niu Y, Ye Y, Zhao W, Shu J. Dynamic monitoring and data analysis of a long-span arch bridge based on high-rate GNSS-RTK measurement combining CF-CEEMD method. *J Civ Struct Health Monit*. 2021;11:35-48.
7. Wang J, Zhao J, Liu Y, Shan J. Vision-based displacement and joint rotation tracking of frame structure using feature mix with single consumer-grade camera. *Struct Control Health Monit*. 2021;28(12):e2832.
8. Cao S, Yan J, Nian H, Xu C. Full-field out-of-plane vibration displacement acquisition based on speckle-projection digital image correlation and its application in damage localization. *Int J Mech Syst Dyn*. 2022;2(4):363-373.
9. Yu Q, Yin Y, Zhang Y, Chen W, Hu B, Liu X. Displacement measurement of large structures using nonoverlapping field of view multi-camera systems under six degrees of freedom ego-motion. *Comput-Aided Civ Infrastruct Eng*. 2023;38(11):1483-1503.
10. Wahbeh AM, Caffrey JP, Masri SF. A vision-based approach for the direct measurement of displacements in vibrating systems. *Smart Mater Struct*. 2003;12(5):785-794.
11. Horn BKP, Schunck BG. Determining optical flow. *Artif Intell*. 1981;17(1-3):185-203.
12. Shi J. Good features to track. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 1994:593-600.
13. Chu TC, Ranson WF, Sutton MA. Applications of digital-image-correlation techniques to experimental mechanics. *Exp Mech*. 1985;25:232-244.
14. Lowe DG. Object recognition from local scale-invariant features. In: *Proceedings of the seventh IEEE International Conference on Computer Vision*. Vol 2. IEEE Computer Society; 1999:1150-1157.
15. Khuc T, Nguyen TA, Dao H, Catbas FN. Swaying displacement measurement for structural monitoring using computer vision and an unmanned aerial vehicle. *Measurement*. 2020;159:107769.
16. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF). *Comput Vis Image Underst*. 2008;110(3):346-359.
17. Kalal Z, Mikolajczyk K, Matas J. Forward-backward error: automatic detection of tracking failures. In: *2010 20th International Conference on Pattern Recognition*. IEEE Computer Society; 2010:2756-2759.
18. Dong CZ, Catbas FN. A non-target structural displacement measurement method using advanced feature matching strategy. *Adv Struct Eng*. 2019;22(16):3461-3472.
19. Bolme DS, Beveridge JR, Draper BA, Lui YM. Visual object tracking using adaptive correlation filters. In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2010:2544-2550.

20. Henriques JF, Caseiro R, Martins P, Batista J. High-speed tracking with kernelized correlation filters. *IEEE Trans Pattern Anal Mach Intell*. 2015;37(3):583-596.
21. Zuo W, Wu X, Lin L, Zhang L, Yang MH. Learning support correlation filters for visual tracking. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(5):1158-1172.
22. Zhao J, Bao Y, Guan Z, Zuo W, Li J, Li H. Video-based multiscale identification approach for tower vibration of a cable-stayed bridge model under earthquake ground motions. *Struct Control Health Monit*. 2019;26(3):e2314.
23. Helfrick MN, Niezrecki C, Avitabile P, Schmidt T. 3D digital image correlation methods for full-field vibration measurement. *Mech Syst Signal Process*. 2011;25(3):917-927.
24. Baqersad J, Niezrecki C, Avitabile P. Extracting full-field dynamic strain on a wind turbine rotor subjected to arbitrary excitations using 3D point tracking and a modal expansion technique. *J Sound Vib*. 2015;352:16-29.
25. Almeida G, Melício F, Biscaia H, Chastre C, Fonseca JM. In-plane displacement and strain image analysis. *Comput-Aided Civ Infrastruct Eng*. 2016;31(4):292-304.
26. Zhang C, Elaksher A. An unmanned aerial vehicle-based imaging system for 3D measurement of unpaved road surface distresses. *Comput-Aided Civ Infrastruct Eng*. 2012;27(2):118-129.
27. Tian Y, Zhang C, Jiang S, Zhang J, Duan W. Noncontact cable force estimation with unmanned aerial vehicle and computer vision. *Comput-Aided Civ Infrastruct Eng*. 2021;36(1):73-88.
28. Park HS, Lee HM, Adeli H, Lee I. A new approach for health monitoring of structures: terrestrial laser scanning. *Comput-Aided Civ Infrastruct Eng*. 2007;22(1):19-30.
29. Park K, Torbol M, Kim S. Vision-based natural frequency identification using laser speckle imaging and parallel computing. *Comput-Aided Civ Infrastruct Eng*. 2018;33(1):51-63.
30. Zhao J, Hu F, Xu Y, Zuo W, Zhong J, Li H. Structure-PoseNet for identification of dense dynamic displacement and three-dimensional poses of structures using a monocular camera. *Comput-Aided Civ Infrastruct Eng*. 2022;37(6):704-725.
31. Dosovitskiy A, Fischer P, Ilg E, et al. FlowNet: learning optical flow with convolutional networks. In: *Proceedings of the IEEE International Conference on Computer Vision*. IEEE Computer Society; 2015:2758-2766.
32. Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, Brox T. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society; 2017: 2462-2470.
33. Choi LK, Bovik AC. Video quality assessment accounting for temporal visual masking of local flicker. *Signal Process: Image Commun*. 2018;67:182-198.

How to cite this article: Zhao J, Li H, Xu Y. Nodes2STRNet for structural dense displacement recognition by deformable mesh model and motion representation. *Int J Mech Syst Dyn*. 2023;3:229-250. doi:10.1002/msd2.12083

APPENDIX

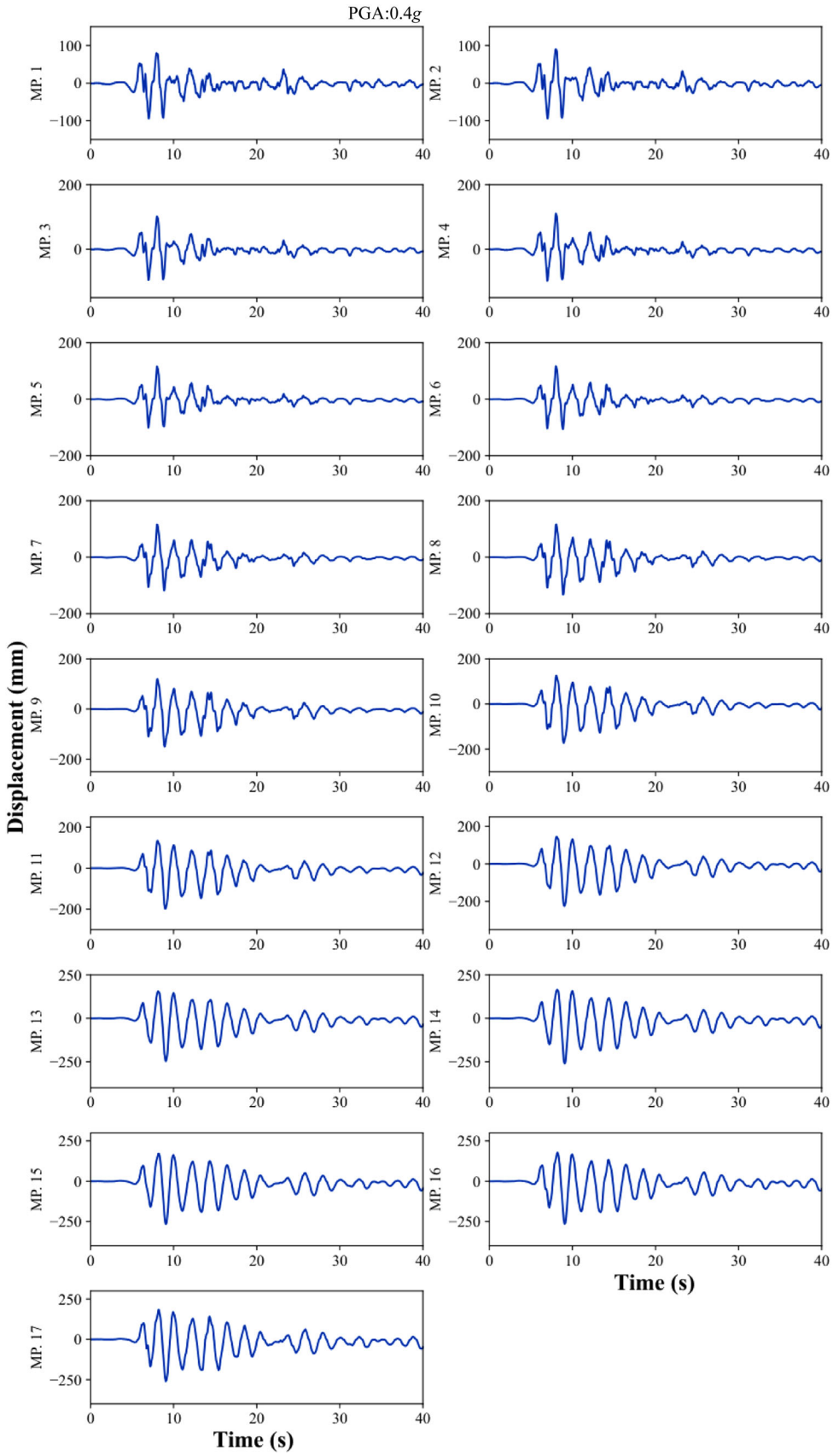

FIGURE A1 Dense displacement recognition results by Nodes2STRNet in BM18. PGA, peak ground acceleration.

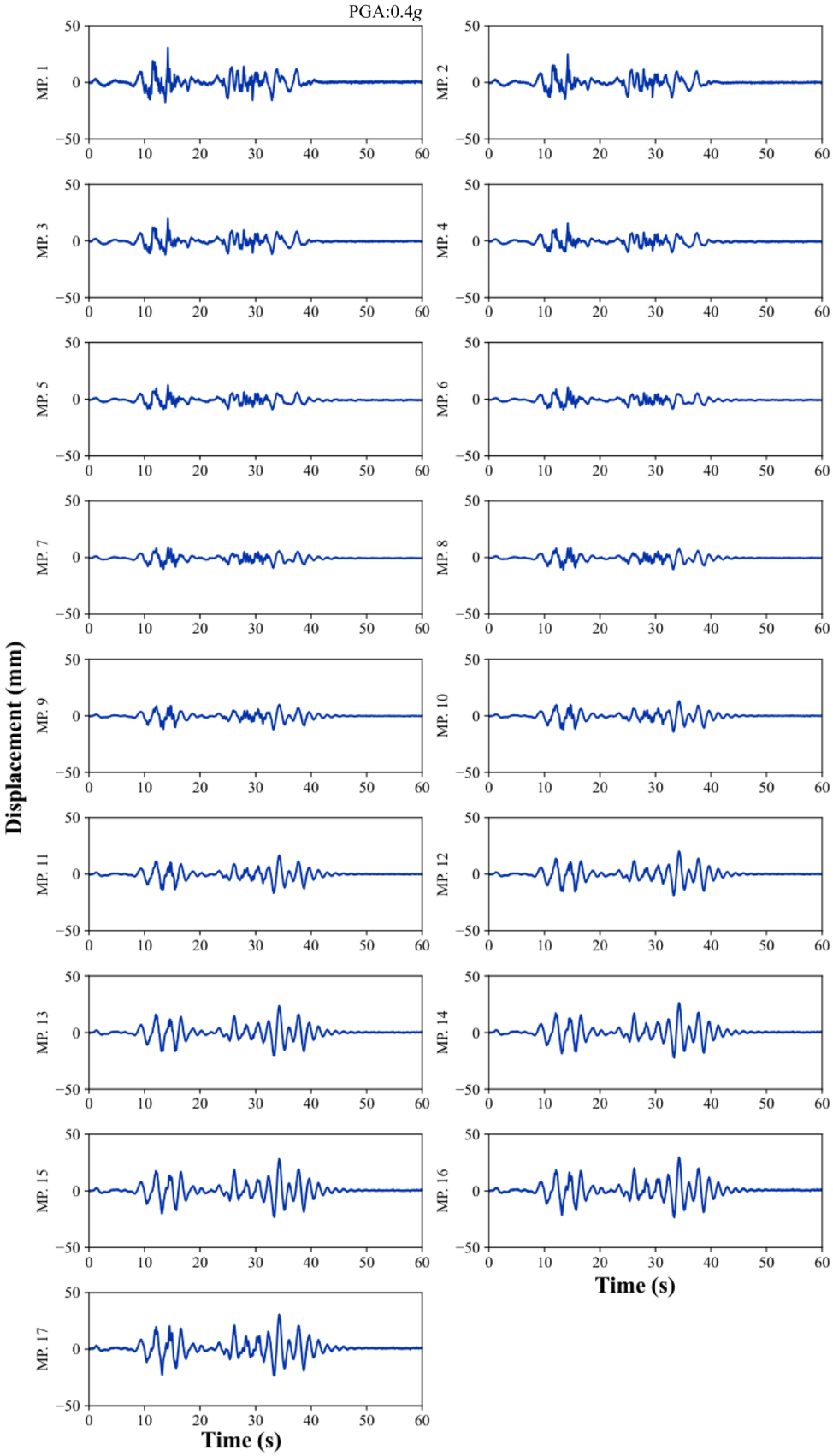

FIGURE A2 Dense displacement recognition results by Nodes2STRNet in BM19. PGA, peak ground acceleration.

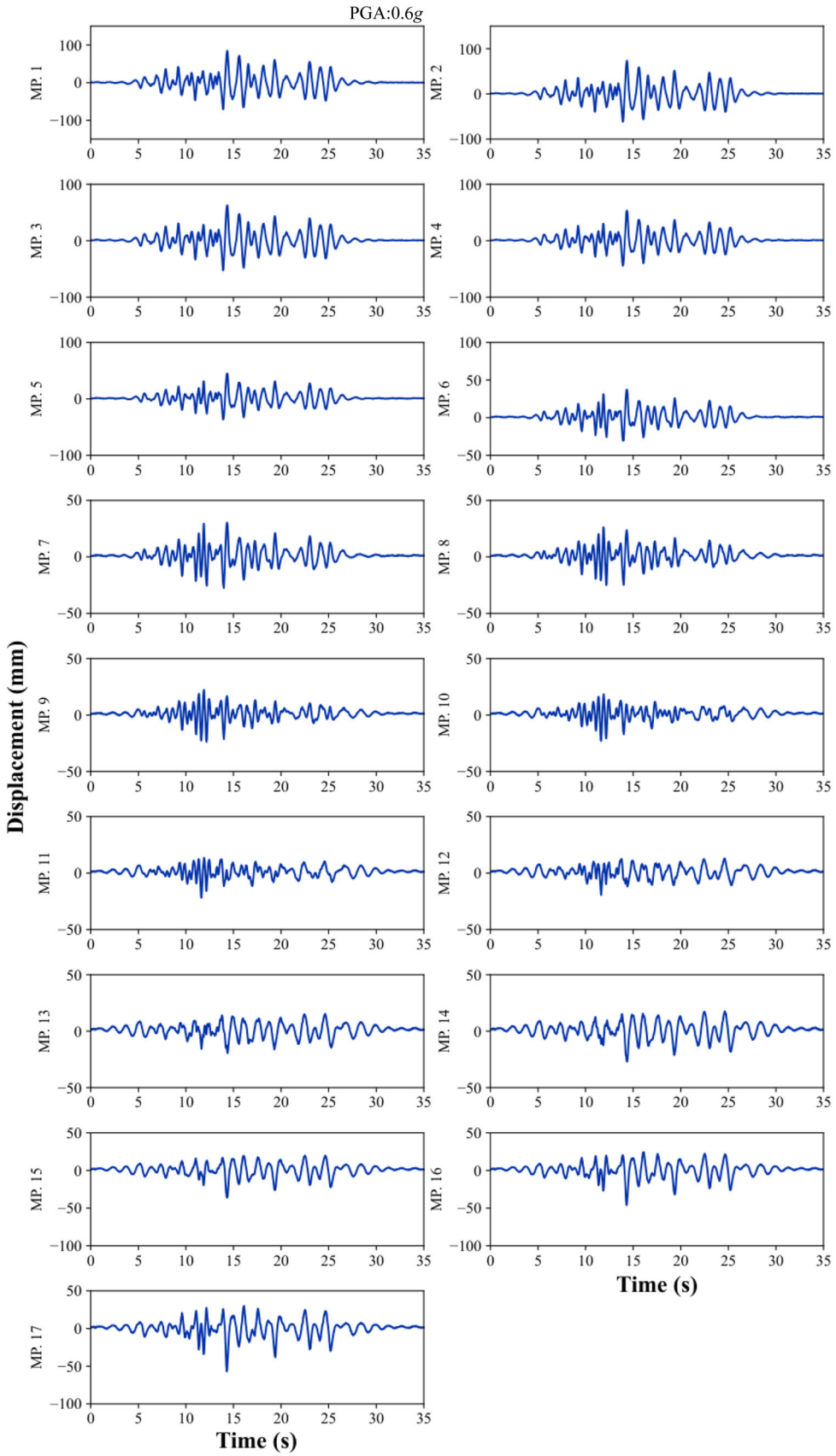

FIGURE A3 Dense displacement recognition results by Nodes2STRNet in BM22. PGA, peak ground acceleration.

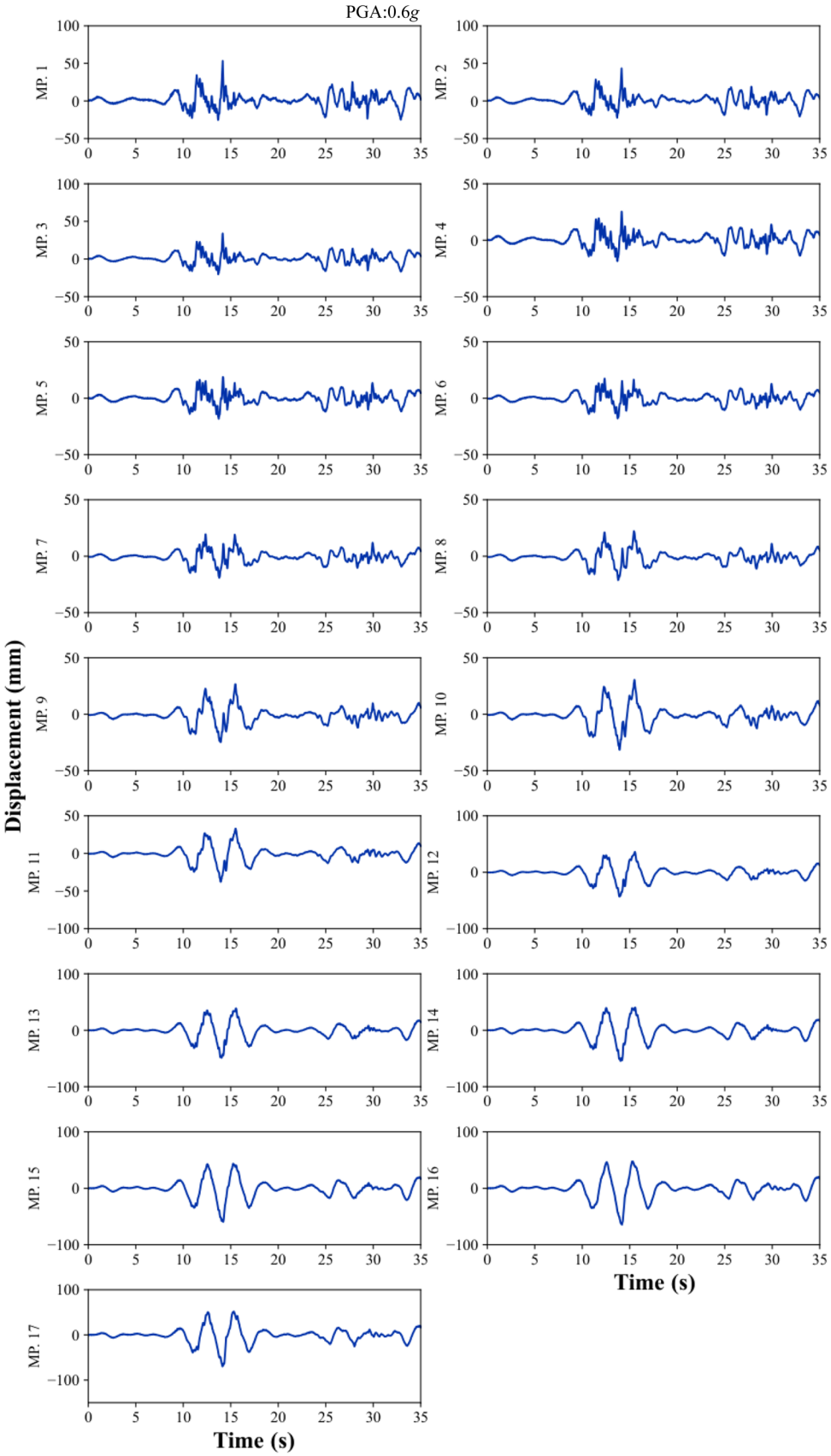

FIGURE A4 Dense displacement recognition results by Nodes2STRNet for BM25. PGA, peak ground acceleration.

AUTHOR BIOGRAPHIES

Jin Zhao received his Ph.D. degree in civil engineering from the Harbin Institute of Technology in 2022. His research interests include structural health monitoring and computer vision.

Hui Li is currently a professor at the Laboratory of Artificial Intelligence and School of Civil Engineering at the Harbin Institute of Technology. She is the founder director of the Key Lab of Smart Prevention and Mitigation of Civil Engineering Disasters of the Ministry of Industry and Information Technology. Her research interests include machine learning for science. She has authored over 200 journal papers, including research articles in Science and Nature. She was awarded the ASCE George W. Housner Medal (2021), ASCE Robert H. Scanlan Medal (2023), and the Structural Health Monitoring Person of the Year Award (Stanford, 2015). She was the ex-president of the International Association for Structural Control and Monitoring and is the current president of the Asian Pacific Network of Centers for Research in Smart Structure Technology.

Yang Xu received his B.S., M.S., and Ph.D. degrees in civil engineering and mechanics from the Harbin Institute of Technology in 2012, 2014, and 2019, respectively. He was a visiting scholar at the University of California, Berkeley from 2017 to 2018. He is currently an associate professor at the School of Civil Engineering at the Harbin Institute of Technology. His research interests include structural health monitoring, computer vision, and deep learning. He was awarded the China National Postdoctoral Program for Innovative Talents.