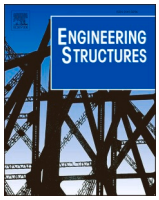
Engineering Structures 318 (2024) 118767

Contents lists available at ScienceDirect

Engineering Structures

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



3D displacement measurement using a single-camera and mesh deformation neural network

Yanda Shao [a], Ling Li [a,*], Jun Li [b,*], Qilin Li [a], Senjian An [a], Hong Hao [c,b]

[a] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, Western Australia 6102, Australia
[b] Centre for Infrastructural Monitoring and Protection, School of Civil and Mechanical Engineering, Curtin University, Bentley, Western Australia 6102, Australia
[c] Earthquake Engineering Research and Test Centre, Guangzhou University, Guangzhou, China

ARTICLE INFO

Keywords:
Monocular Vision
Vibration Displacement Measurement
Structural Health Monitoring
Deep Learning

ABSTRACT

Computer vision-based methods for civil structure's vibration displacement measurement have emerged as useful tools in the recent years. These methods offer several benefits including non-contact measurements, cost-effectiveness, and the ability to capture full-field displacement. Yet, there remain certain challenges. Measuring vibration displacement in 3D typically requires multiple cameras, adding complexity to camera configurations. Moreover, existing methods relied heavily on physical markers or natural key points. Placing physical markers on structures is often impractical, and natural key points are difficult to detect on structures with few distinct features or during rapid movements. Contrary to previous approaches, this paper presents a novel technique that uses a monocular camera for 3D displacement measurements. This technique obviates the need for physical markers or the reliance on natural key points, representing a significant advancement. Central to the method is a deep neural network designed to predict 3D mesh deformation directly from a single image input, combined with an initial 3D cube mesh input. A synthetic 3D dataset is generated to train the neural network. In the testing phase for real structures, advanced video segmentation method is employed to remove the background in order to enhance the prediction accuracy. The practical efficacy of this methodology is validated in a laboratory through a series of experimental tests on beam structures, demonstrating reliable results and application potentials.

1. Introduction

Structural vibration displacement serves as an important indicator of structural performance therefore can be used for civil structural health monitoring (SHM) [1–3]. By analysing the vibrations, engineers can extract the structural integrity and potential vulnerabilities inherent within the system. When a structure is subjected to external forces, such as vehicular traffic on a bridge or wind forces on a wind turbine, it undergoes certain displacements. Anomalies in the expected displacement patterns can be indicative of structural deficiencies or material fatigue [4,5]. Continuous monitoring of these displacements facilitates a proactive approach to structural maintenance. It allows for the early detection of potential issues, enabling timely interventions. This ensures the safety and life of the structure and optimizes the economic aspects by preventing costly repairs. Post-intervention evaluations are equally crucial. By comparing displacement data before and after repair or retrofit interventions, engineers can quantitatively assess the efficacy of the measures undertaken. If the post-repair displacements align more closely with the expected patterns, it indicates a successful intervention. Conversely, continued anomalies might necessitate further investigative measures or additional corrective actions.

Traditional methods for displacement measurement typically involve the use of contact sensors such as linear variable differential transformers (LVDTs), strain gauges and so on. These sensors are attached to the surface of the structure to measure the movement or deformation of the structure under load. While traditional methods can provide accurate measurements, they are often limited to localized measurements, requiring multiple sensors to monitor the behaviour of the entire structure. For example, LVDTs can only measure the movement or deformation of a particular point on the structure. This limitation can result in a partial understanding of the structural behaviour, leading to missed opportunities for early detection of damage or

* Corresponding authors.
E-mail addresses: yanda.shao@postgrad.curtin.edu.au (Y. Shao), L.Li@curtin.edu.au (L. Li), junli@curtin.edu.au (J. Li), Qilin.li@curtin.edu.au (Q. Li), S.An@curtin.edu.au (S. An), hong.hao@curtin.edu.au (H. Hao).

https://doi.org/10.1016/j.engstruct.2024.118767
Received 18 January 2024; Received in revised form 5 June 2024; Accepted 8 August 2024
Available online 15 August 2024

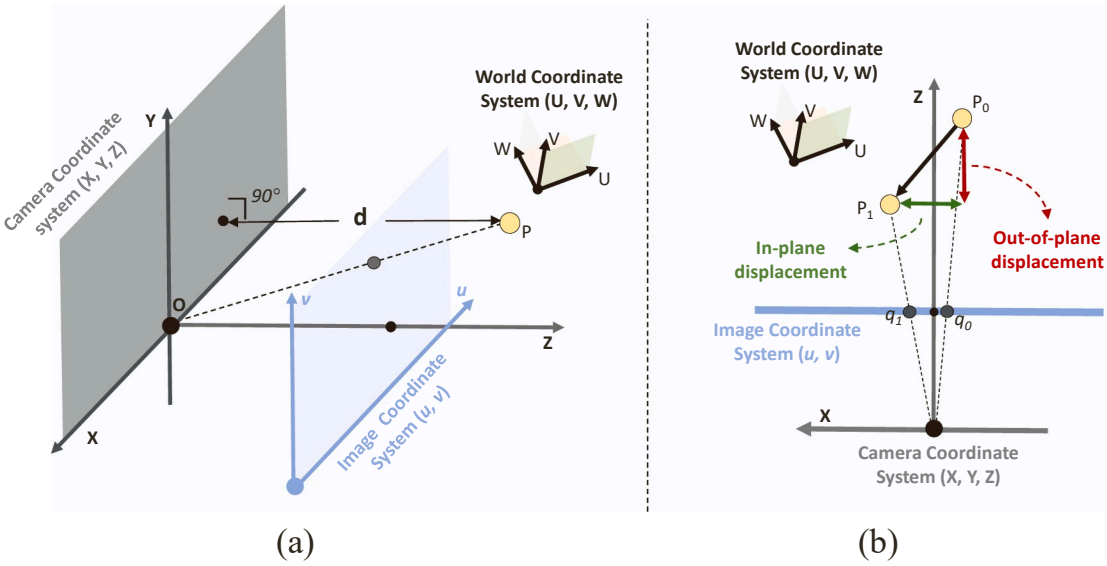


Fig. 1. Depiction of in-plane and out-of-plane displacements in relation to the pinhole camera model. (a) Emphasizes the pinhole camera model, highlighting the inherent limitation in capture depth information; (b) Demonstrates both in-plane and out-of-plane movements of a specific point.

deformation. Additionally, these methods may require invasive installation procedures, such as drilling holes or attaching sensors to the structure, which can damage the structural integrity. The installation of conventional sensors can be time-consuming, labour-intensive, and may necessitate temporary disruptions to the structure's function or use [6]. Moreover, the maintenance and calibration of these sensors over time can add to the long-term costs and logistical challenges of structural monitoring.

There has been a significant shift towards non-contact sensors, which offer the advantage of measuring displacements without physically touching the target. Laser displacement sensors [7] are among the most precise non-contact measurements. However, their effective range is limited, and require a static platform for installation to ensure consistent and accurate measurements. Satellite-based remote sensing techniques can also be used to measure displacement of structures. The common technique is Global Navigation Satellite System (GNSS) [8–10], which uses signals from satellite navigation systems to measure displacement of structures. This technique is particularly useful for monitoring large structures such as bridges or dams that may be difficult or dangerous to access for installation of traditional sensors for displacement measurement. Satellite-based remote sensing techniques have limitations including low resolution due to the size of the satellite sensor and distance from the ground, as well as environmental factors such as atmospheric conditions and vegetation that would affect the measurements.

Computer vision-based displacement measurement techniques offer several advantages over traditional monitoring techniques, as they can provide full-field measurements, capturing the deformation of the entire surface of a structure rather than just sparse points [11]. Dense measurements can provide a more complete picture of the structural behaviour and can aid in the detection of damage, fatigue, or deformation. Being a form of remote sensing, computer vision techniques are especially beneficial for monitoring structures that are challenging to access or where the installation of conventional sensors might be impractical. Furthermore, these methods are relatively easy to set up, eliminating the need for intricate installations or specialized equipment. Additionally, their cost-effectiveness, due to reduced equipment and maintenance expenses, makes them an attractive option for structural health monitoring.

Over the past two decades, computer vision-based displacement measurement systems have undergone significant development. As these systems have evolved, there has been a notable improvement in both of their application scenarios and measurement accuracy. The current computer vision-based measurement systems can be categorized into two classes: 2D displacement systems [1,12–17] and 3D displacement systems [11,18–22]. As shown in Fig. 1(a), the camera imaging process inherently reduces dimensions, projecting 3D objects from the world coordinate system onto a 2D image coordinate system. In such a transformation, the depth information *d*, is lost. As illustrated in Fig. 1 (b), 2D displacement measurement systems are designed to capture only the in-plane displacements that occur parallel or approximately parallel to the image plane. In contrast, 3D displacement measurement attempts to capture additionally the out-of-plane movement, which are perpendicular to the camera's imaging plane. Measuring out-of-plane displacement presents significantly more challenges compared to in-plane displacement measurement, primarily due to the absence of depth information in images.

Due to the inherent characteristics of various structures and the diverse loadings they are exposed to, out-of-plane displacements are often unavoidable. Firstly, structures frequently experience dynamic loads, such as wind, seismic activity, or vehicular traffic. These forces can induce not only in-plane movements but also complex three-dimensional displacements. Secondly, even when a structure's movement is primarily one or two-dimensional, it is often challenging to achieve a camera position perfectly so the imaging place is parallel to the movement. Environmental constraints, installation hurdles, or the necessity to view multiple aspects of a structure can hinder ideal camera placement.

Over the past decades, many measurement systems have been designed to measure the out-of-plane displacement, many harnessing the capabilities of depth sensors [23] or relying on multi-camera arrays [19,24] for accurate measurements. The depth sensors are commendable in their efficacy, but are accompanied by inherent challenges. For example, they are susceptible to environmental variables. Ambient lighting, shadows, and reflections can introduce anomalies in their readings. The cost factor further compounds the issue; high-quality depth sensors often come with a significant price tag. Given these challenges, multi-camera measurement systems have gained popularity [11,19–21,24]. These systems leverage multiple cameras to capture different perspectives of a structure. By leveraging images from various viewpoints, they can reconstruct a 3D representation of the structure's movements, recovering the lost depth information in single-camera setups. Park et al. [19] introduced an approach that explores the utilization of a multi-camera based motion capture system (MCS) for capturing 3D displacements of structures. Their findings underscore the

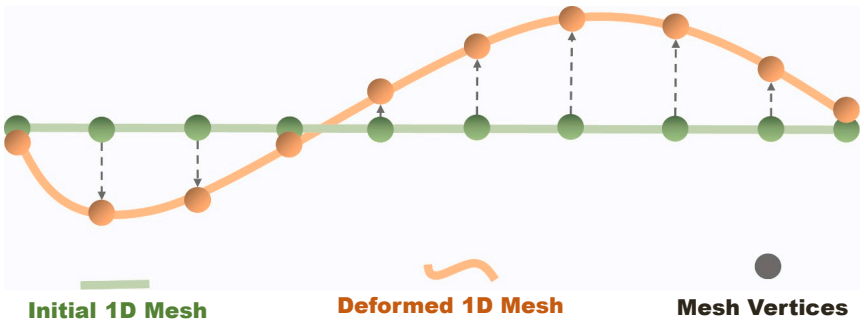


Fig. 2. An example of 1D mesh deformation. Each circle represents a vertex.

potential of MCS in offering a more comprehensive recording of structural behaviours, especially in scenarios where torsional and lateral movements coexist. Shao et al. [11] introduced a computer vision-based approach for full-field 3D vibration displacement measurement without the need for artificial targets. Utilizing a binocular vision system, the method captures both in-plane and out-of-plane vibration displacements in civil structures. A significant advancement of this study is the integration of deep learning-based key point detection and matching algorithms, enabling highly precise target-free measurements. Wang et al. [21] introduced a binocular vision measurement system for structural health monitoring, leveraging the SIFT-based improved LSM algorithm (SLSM). This system, equipped with a 4 G cellular module, allows for real-time image uploading and processing. However, multi-camera setups pose their own set of challenges in real applications. Multi-camera configurations demand rigorous calibration to achieve uniform depth perception from diverse angles. Techniques like key point matching, while effective, are computationally demanding and intricate, amplifying the system's complexity. Additionally, synchronizing multiple cameras is also a difficult task.

Recently, the use of monocular camera systems for out-of-plane displacement measurement has emerged as a promising technique, offering several advantages over multi-camera configurations. At its core, the monocular approach, relying on just a single camera, stands out for its simplicity and cost-effectiveness. There is no need for the complexities of multi-camera installations, or the intricate calibration processes that come with ensuring multiple cameras synchronize and align correctly. The recent strides in deep learning have significantly bolstered this approach. Contemporary research, empowered by deep learning advancements, has demonstrated the potential of monocular cameras in gauging out-of-plane movements [25–27]. In recent work, Sun et al. [27] presented an approach that harnesses monocular vision combined with deep learning-based pose estimation to effectively measure 3D displacements using just a single camera. Leveraging virtual rendering, the authors adeptly created extensive training datasets derived from 3D models of target structures, thereby circumventing the labour-intensive task of manual annotations. While the method showcases significant potential for accurate rigid body displacement measurements, it exhibits limitations in detecting displacements when the structure undergoes deformation, suggesting further enhancements are required to tackle non-rigid body dynamics. Shao et al. [26] introduced a measurement system that leverages the capabilities of a monocular vision system for 3D full-field displacement measurement in civil structures. The system combines advanced key point detection and tracking algorithms for in-plane displacement measurement with advanced deep learning techniques for out-of-plane depth estimation. This fusion allows for the measurement of 3D displacement using just a single stationary camera. The experimental results highlighted the system's capability in accurately capturing the in-plane vibration displacement responses. However, the accuracy of out-of-plane measurements, while commendable, is far from comparable to the in-plane measurements.

The measurement for both in-plane and out-of-plane displacement can be divided into target-based and target-free approaches. The former requires the placement of artificial markers on the structure, which simplifies the tracking process for computer vision algorithms and enhances accuracy. However, the installation of these markers can be labour-intensive and time-consuming, and their presence limits the full-field displacement measurements. On the other hand, target-free methods capitalize on the natural features of the structure or the patterns of ambient light and shadow to ascertain displacement. With the integration of advanced algorithms, such as SIFT [28], SURF [29], and KAZE [30], and the incorporation of deep learning techniques like Superpoint [31], these methods can extract and track natural features over time to determine movement. However, structures that lack distinct textures can pose significant difficulties for target-free measurement, as the absence of unique visual features can hinder accurate tracking and measurement. Similarly, structures undergoing rapid movements can introduce motion blur or exceed the tracking capabilities, leading to potential inaccuracies.

This paper proposes a method centering on the concept of mesh deformation, obviating the need for key point detection. Essentially, a mesh serves as a skeletal framework of an object, constructed from vertices, edges, and faces that collectively represent the object's three-dimensional form. Mesh deformation involves manipulating these vertices, much like moulding clay, to allow for alterations in the object's shape. Fig. 2 shows a simple example of a 1D mesh structure composed of ten vertices. Initially, the mesh is presented as a straight line. When the positions of these vertices are altered, deformation of the mesh is observed. The displacement of the vertices from the initial to the deformed mesh can be measured. The greater the number of vertices in a mesh is, the more comprehensive the full-field displacement can be captured.

Analogous to children moulding clay based on a reference picture, this study utilizes an initial mesh cube as a simple yet malleable 3D form, akin to a piece of unshaped clay. This analogy serves to illustrate the adaptability of the mesh deformation approach, where the initial mesh, like clay, is flexible and can be shaped or adjusted. The mesh deformable neural network enhances this malleability, allowing it to modify the vertices of the mesh in 3D based on a single reference image. This technique ensures that the mesh adapts holistically to structural movements, rather than focusing solely on isolated key points. Each vertex in the mesh correlates to a specific point on the structure, and as the structure moves or displaces, its representation in the image shifts accordingly. The altered image informs the neural network to adjust the vertices' positions, effectively capturing the mesh deformation that reflects the structural changes.

This vertex-focused approach, which prioritizes the collective deformation of mesh vertices over isolated key point tracking, offers a robust and holistic framework to structural displacement measurement. This is especially advantageous in scenarios where traditional methods might fail. To train a neural network capable of inferring mesh deformation from a single image, a large synthetic dataset is generated. This

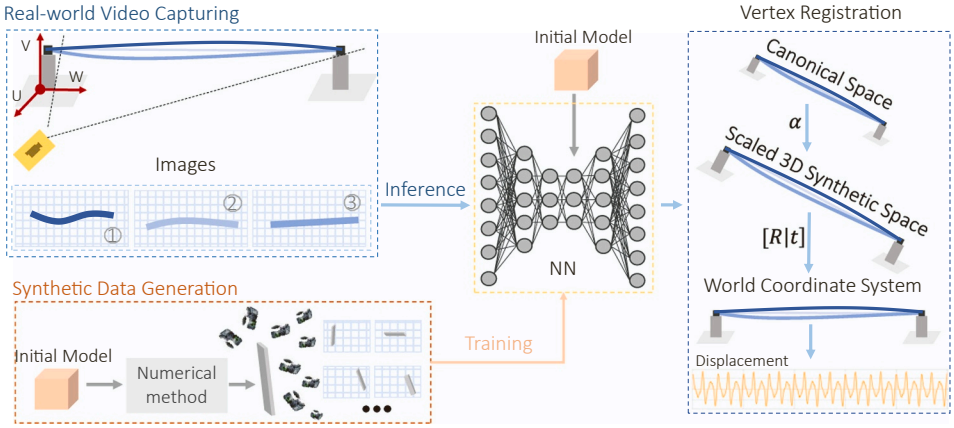


Fig. 3. The overview of the proposed displacement measurement system.

dataset encompasses a wide range of deformed meshes and corresponding images captured from various viewpoints. To eliminate the effect of background in testing, a robust vision model Track Anything (TAM) [32] is utilized to segment the desired structure from each frame of the structural vibration video. By integrating TAM, a significant improvement in measurement accuracy is observed.

The proposed approach offers three advantages over traditional methods. Firstly, the method is characterized by its **1) target-free** nature, eliminating the need for key point tracking or the placement of physical markers on the structure. This not only streamlines the monitoring process but also reduces potential errors that might arise from marker placement or tracking discrepancies. Furthermore, **2) out-of-plane displacements** are captured using monocular vision, a significant advancement from conventional techniques that often require multiple cameras or sensors. **3) A full-field view** of structural movements is provided by emphasizing mesh deformation over individual key points, ensuring a comprehensive understanding of structural dynamics. This is especially pivotal for structures that cannot be adequately represented by isolated key points or when the movement is not indicative of rigid body motion.

The remainder of this paper is structured as follows: Section 2 introduces the generation of the dataset, the intricacies of the mesh deformation neural network, and an introduction to TAM. The efficacy of the proposed vision measurement system is then assessed in Section 3 through a series of vibration tests conducted on beam structures. Section 4 provides a comprehensive discussion of the proposed measurement methodology. Conclusions are drawn in Section 5.

2. Methodology

In the development of this 3D displacement measurement system, a systematic and multi-step process is followed, as illustrated in Fig. 3. At the outset, a custom-made dataset is generated using the 3DGEN [33] synthetic dataset generation tool. The generated dataset serves as the training data for the mesh deformation neural network model. This model can predict the deformations of the 3D mesh based on a single image. Before testing the model, a crucial preprocessing step is carried out using TAM [32] to eliminate the background from the test video. TAM's primary function is to enhance the videos by isolating the object of interest from its background. This segmentation process involves the elimination of all background elements, a necessary step given the absence of background in the training dataset and the challenges associated with generating a diverse background within the dataset. Eliminating the background ensures that the subsequent analysis remains unpolluted by any irrelevant objects. The outcome of this integration is a series of deformed 3D meshes. An assumption of this methodology is the existence of at least one stationary point within the civil structure. This assumption is not only logical but also practical, as most civil structures inherently possess a fixed foundation. By measuring the distance between the stationary vertex (vertices) and the dynamic ones within each deformed mesh, the displacement can be measured. Since the mesh deformation network is trained on images from various viewing angles, it allows the flexibility of positioning the camera at any desired place, ensuring accurate out-of-plane displacement measurements.

2.1. 3D mesh deformation neural network

The deformation of the 3D mesh relies on a combination of deep learning and numerical methods. Convolutional neural network (CNNs) is used to translating 2D changes into 3D meshes. Alongside deep learning, analytical methods are used to calculate the deformations of structures under various loading conditions, guiding the training of the

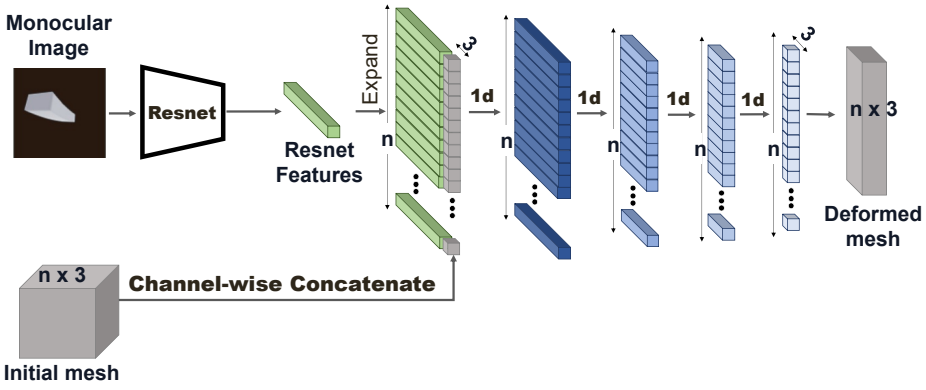


Fig. 4. The architecture of the neural network for 3D mesh deformation.

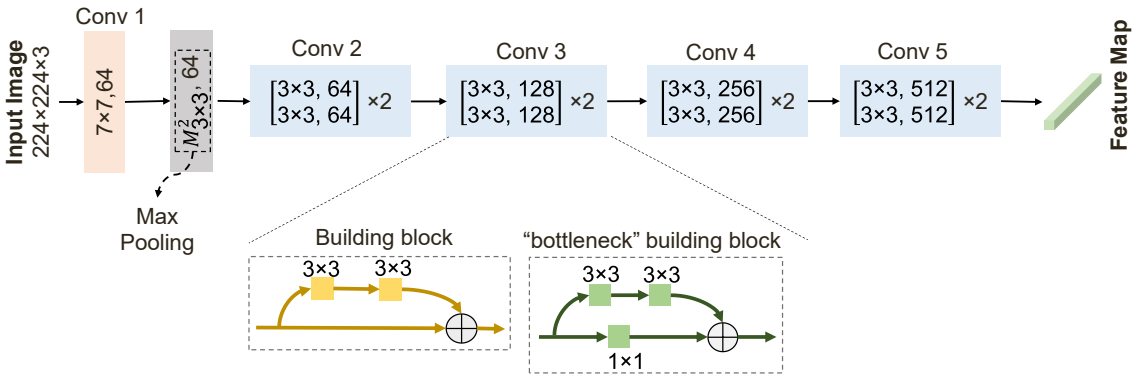


Fig. 5. Schematic representation of the ResNet-18 architecture: illustrating the layer-wise composition and residual connections.

neural network.

As shown in Fig. 4, the mesh deformation neural network operates using an encoder-decoder architecture. This network takes in a single monocular image and an initial mesh, then produces a deformed mesh corresponding to the input. For the feature extraction phase, residual neural network (ResNet) [34] serves as the encoder, effectively capturing the significant details and patterns in the input image. Subsequently, 1D CNNs are employed as the decoder, reconstructing the deformed mesh from the feature map extracted by ResNet. This combination allows the network to adaptively shape the mesh based on the intricacies and details perceived in the monocular image.

The encoder serves as a foundational component in various neural network architectures. Its primary role is to compress and transform raw input data into a more compact and informative representation. The encoder of the mesh deformation neural network takes in a monocular RGB image as input and extracts high-level features from the image using a series of convolutional layers. ResNet-18 has been selected as the backbone for feature extraction. This specific variant of the ResNet family, with its optimized depth, is characterized by a balance between computational efficiency and the ability to identify intricate patterns in the data.

The ResNet-18 architecture is delineated with 18 layers, of which 17 are convolutional and one stands as a fully connected layer. Initially the structure is a 7×7 convolutional layer featuring a stride of 2, promptly succeeded by a max pooling layer. Subsequently, the 16 layers evolve into a series of residual blocks. Each block encompasses two 3×3 convolutional layers, integrated with a shortcut connection capable of sidestepping one or several layers. A distinguishing feature of this architecture is its residual nature, facilitating a more streamlined training process and elevating performance in deep neural networks. Every convolutional layer in ResNet-18 is followed by batch normalization, a strategy that greatly enhances the stability and accelerates the training phase. Before reaching the final fully connected layer, a global average pooling layer is incorporated, effectively diminishing the number of parameters and mitigating the risk of overfitting. A comprehensive layout of the ResNet-18 structure can be referenced in Fig. 5.

The decoder of the model is a 1D CNN that takes the combined feature representation from the encoder and the vertices of the original mesh as input and generates deformed vertices. In the proposed neural network model, the initial operation begins with the output from the encoder. Subsequently, the output from the encoder is expanded along a new dimension to match the size of the original mesh, which contains the 3D coordinates of the vertices. This expansion effectively replicates the encoder's output for each point in the original mesh. Finally, the expanded encoder output is concatenated with the vertex tensor along the feature dimension. This concatenation operation combines the image-derived features with the 3D point data, creating a comprehensive feature representation. The decoder is defined with four convolutional layers and three batch normalization layers. The first three layers use the Rectified Linear Unit (ReLU) [35] activate function, while the last output layer uses the hyperbolic tangent (tanh) activate function.

During training, three loss functions are used: Chamfer loss, edge length regularization loss and a surface normal smooth loss [36]. Chamfer loss measures the average proximity of each vertex in one set (prediction) to its nearest neighbour in the other set (ground truth). The loss is bidirectional: it calculates the distance from every point in the first set to its closest point in the second set and vice versa, then combines these distances to produce a comprehensive measure of point set dissimilarity. The core advantage of the Chamfer loss lies in its ability to provide a robust distance metric even when two-point sets have different cardinalities or are sparsely sampled. The formula of the Chamfer loss is provided as:

$$l_{CD}(P,Q) = \frac{1}{|p|}\sum_{p\in P}\min_{q\in Q}\|p-q\|_2^2 + \frac{1}{|q|}\sum_{q\in Q}\min_{p\in P}\|p-q\|_2^2. \tag{1}$$

where P and Q represent the group of predicted vertices and ground truth vertices. The first term in the equation calculates the average minimum squared distance from each point in P to Q, and the second term does the opposite, from each point in Q to P. In the domain of 3D mesh analysis, accurately capturing the underlying structural characteristics of mesh is important.

To penalize flying vertices, an edge length regularization loss [36] is added:

$$\mathrm{l}_{edge} = \sum_{p}\sum_{k\in\mathcal{N}(p)}\|p-k\|_2^2, \tag{2}$$

where p for a vertex in the predicted mesh, k for a neighbouring vertex of p. The edge length regularization loss is a quantitative measure that assesses the variations in the lengths of edges formed by pairs of adjacent vertices in a 3D mesh. Utilizing the edge loss in optimization tasks serves a dual purpose: it not only aids in producing geometrically consistent models but also ensures that the resultant structures adhere closely to the intricate nuances of the target data.

For an enhanced characterization of high-order geometric properties, a surface normal loss term [36] is added as:

$$\mathrm{l}_n = \sum_{p}\sum_{q=argmin(\|p-q\|_2^2)}\left\|\langle p-k, n_q\rangle\right\|_2^2, \quad s.t. \quad k\in\mathcal{N}(p). \tag{3}$$

where p and q are corresponding vertices in P and Q, found during the computation of a preceding distance-based loss (such as Chamfer loss). The term k represents a neighboring vertex of p and $\mathcal{N}(p)$ is the set of such neighbors. n_q is the observed surface normal from the ground truth. The inner product between two vectors is denoted as $\langle \bullet, \bullet \rangle$. The overall loss is a weighted sum of two losses, $loss_{overall} = l_{CD} + 0.05l_{edge} + 0.025\mathrm{l}_n$.

The Adam optimizer [37], utilizing its default parameters, is used to update the network's weights. The initial learning rate is configured at

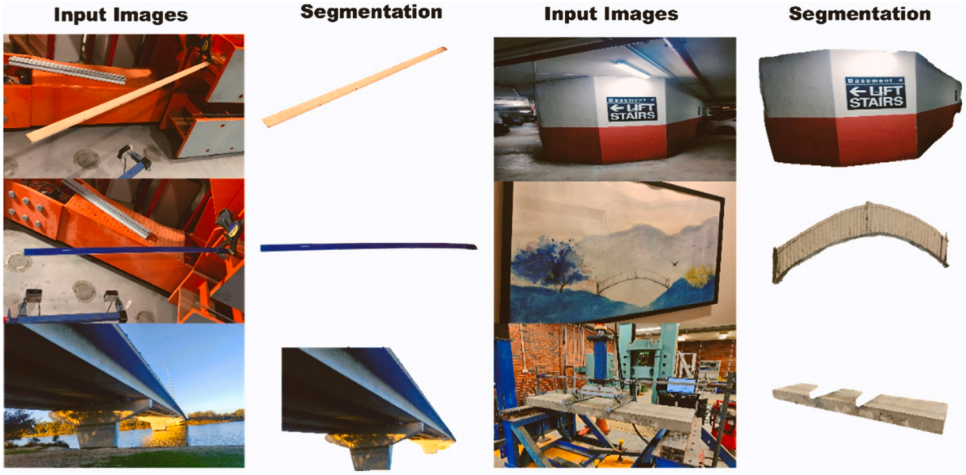


Fig. 6. Sampled segmentation results from the Tracking Anything method.

1×10^{-3}, and adjustments are designed to be made at three pivotal epochs: firstly, at the 15th, subsequently at the 25th, and ultimately at the 35th, culminating in a total of 50 epochs for the entire training duration. At each designated epoch, the learning rate undergoes a decimation by a factor of ten, a strategy orchestrated to ensure nuanced weight adjustments that foster model convergence. The training exploits the computational capabilities of four Nvidia TITAN RTX GPUs. With a batch size of 128, the training process is streamlined, and to further expedite data loading, all mesh files are converted to NumPy files, facilitating swift 3D mesh loading.

2.2. Point registration for metric displacement measurement

In the field of Structural Health Monitoring (SHM), measurement of metric displacement like millimetres, is of paramount importance. The process of point registration is a critical role in this context. Specifically, the objective is to align two disparate sets of points - typically a synthetic mesh predicted by the neural network to the actual coordinates in the world system. This alignment ensures that the spatial locations captured in the synthetic representation correspond to the real-world, thereby facilitating accurate evaluation of structural health metrics.

For 3D vertices registration, an initial alignment is applied to roughly bring the synthetic representation to the real-world coordinate system. This is achieved through Singular Value Decomposition (SVD). As previously noted, civil structures often have fixed portions that remain invariant over time. These static regions offer an advantageous opportunity for point registration, as they can serve as reference points to facilitate the alignment process. While a minimum of three non-collinear points can theoretically define a unique plane in three-dimensional space, a greater number of fixed points is generally advantageous. Utilizing more fixed points improves the constraint on the transformation parameters, thereby enhancing alignment accuracy.

The fixed points on the synthetic mesh are denoted as P, while their corresponding locations in the real-world coordinate system are represented by Q. The centroids C_p and C_q represent the geometric center of the point sets P and Q, respectively. Mathematically, the centroid is defined as the average of all the points in the set.

$$C_p = \frac{1}{N}\sum_{i=1}^{N} p_i \tag{4}$$

$$C_q = \frac{1}{N}\sum_{i=1}^{N} q_i. \tag{5}$$

here N is the number of points in each set, and the sum runs over all these points. This average captures the "center of geometry" of the points, providing a key point around which alignment can be more naturally achieved. The original point sets are transformed into P' and Q' by subtracting their corresponding centroids. This is for SVD-based methods, as it aligns both sets around the origin, facilitating easier computation of rotational and translational transformations.

$$p_i' = p_i - C_p \tag{6}$$

$$q_i' = q_i - C_q. \tag{7}$$

The cross-covariance matrix H is then calculated to find the optimal rotation that aligns the two centered point sets P' and Q'. It is calculated as:

$$H = \sum_{i=1}^{N} p_i' q_i'^{T}. \tag{8}$$

This matrix essentially captures the spatial relationships between corresponding points in two-point sets. Then, SVD decomposes the matrix H into three other matrices U, S, and V:

$$H = USV^T. \tag{9}$$

The columns of U and V are the left-singular and right-singular vectors, and S is a diagonal matrix containing the singular values. These matrices help to extract the optimal rotation for alignment. The rotation matrix R is computed as the product of V and the transpose of U:

$$R = VU^T. \tag{10}$$

In case the determinant of R is negative, which would imply a reflection, it is corrected using:

$$R = V\mathrm{diag}(1, 1, -1)U^T \tag{11}$$

Finally, the translation vector t aligns the centroids of the transformed P and Q:

$$t = C_Q - RC_P \tag{12}$$

After obtaining an initial registration (R, t) through SVD, the Iterative Closest Point (ICP) algorithm [38] is employed to fine-tune this alignment. ICP iteratively minimizes a cost function E, defined as the sum of squared Euclidean distances between each point p_i in P and its closest point q_i in Q. Specifically,

$$E = \sum_{i=1}^{N} \|q_i - (Rp_i + t)\|^2, \tag{13}$$

where R is the rotation matrix and t is the translation vector. The al-

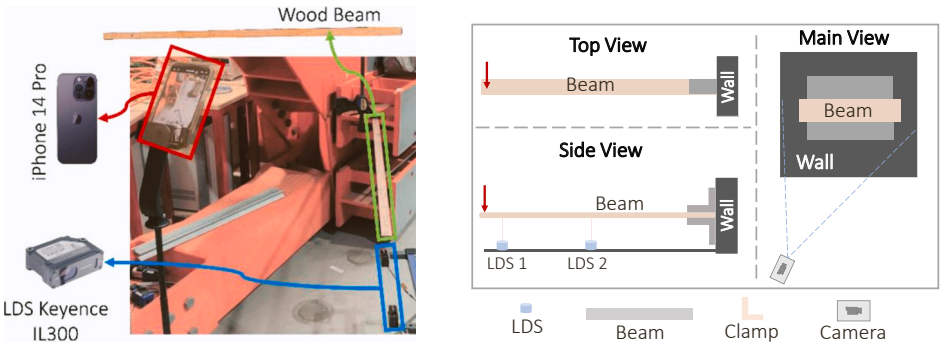


Fig. 7. Experimental Test Setup: (a) Actual photograph showcasing the real-world setup; (b) Schematic diagram illustrating the arrangement and components.

gorithm commences with an initial guess for R and t, provided by the prior SVD-based registration, and refines these parameters until a specified convergence criterion is met. This iterative optimization ensures that P is closely aligned with Q, thus providing a refining point alignment. Once the alignment parameters (R, t) are obtained, they can be applied to the entire synthetic mesh to mapping it to the real-world coordinate system.

2.3. Background removal

Due to the absence of background in the training dataset and the need to minimize distractions during analysis, it is necessary to remove the background in the video frames during the test phase. A large vision model, Tracking Anything [32], is employed to remove the background from each frame of the video. This adaptation ensures that the model's focus remains primarily on the foreground elements, resulting in more precise and targeted analyses.

TAM is a deep learning model for interactive video tracking and segmentation, requiring only minimal user input, such as a handful of clicks on a video sequence. The TAM, begins with an initialization process using Segment Anything [39], enabling users to create a precise mask representation of the chosen object with just a few clicks. This mask is then tracked in subsequent video frames using XMem [40], which conducts semi-supervised Video Object Segmentation (VOS). While XMem predictions are generally satisfactory, instances of suboptimal mask quality are refined using SAM again. This refinement process harnesses parameters from XMem as prompts to produce a more polished segmentation mask. Notably, as the model progresses, challenges in distinguishing objects in intricate scenarios become evident, especially in extended videos. To address this, an option for manual human correction during inference is introduced, empowering users to adjust the mask in the current frame when needed. This integrated approach offers a comprehensive solution for tracking objects in diverse video contexts. Some sampled results of the TAM are shown in Fig. 6. In addition to employing TAM to segment the desired structure from the background, random cropping is implemented during the training phase of the network. Random cropping is used to simulate scenarios where parts of the structure may be obscured or where the boundary lines between the structure and its background are not distinctly clear. By training the model on these modified images, its ability to accurately identify structural features under less-than-ideal conditions is improved, ensuring more reliable displacement measurements across a variety of practical situations.

3. Experimental test for beam structures

3.1. Experiment setup

Beam structure is chosen as the test samples for our experiments since they exist in many complex structures. For example, the structures shown in Fig. 6 can all be decomposed into beam structures of various sizes and analysed by components separately. Two beams are employed for testing: a wooden beam with dimensions of 1200 × 30 × 8 mm and an aluminium beam measuring 1000 × 60 × 3 mm. Each beam is configured as a cantilever, with one end securely anchored to the wall using a clamp to ensure it remains fixed. The schematic of this setup is illustrated in Fig. 7(a), while the on-site experimental setup, including the positioning of the sensors and other components, is illustrated in Fig. 7(b). The free end of the beam is tapped by hand to induce vibrations. For visual recording of the vibrations, a phone camera (iPhone 14Pro) is utilized for filming the experiment. The videos are captured at a resolution of 1920 × 1080 and a frame rate of 155fps (frames per second). The camera in this experiment is pointing towards the movement direction of the object at an unknown angle, so that there always exist out-of-plane movements within the video. Notably, the camera's location varies between the two experiments. This variability in camera positioning is designed to verify that the measurement system can measure displacement from any angle. The system is able to capture the full-field vibration of the whole beam in 3D. For comparison purpose, two LDS Keyence IL300 sensors with a frequency of 200 Hz are installed on the back of the structure to measure the displacement in the vertical direction, providing the ground truth data for our experiments. Without loss of generality, one of these LDSs is positioned at the beam's midpoint, while the other is placed at approximately the ¼ length of the beam from the support.

3.2. Training data generation

The availability of a sufficient dataset is crucial to train a neural network effectively. Due to the lack of real 3D data of civil structures, a synthetic dataset of beam meshes is generated using the customisable data generation system 3DGEN [33]. This dataset consists of two parts: 3D mesh models and the corresponding rendered images from various viewpoints. Base mesh models are crafted to retain a length, width, and height ratio consistent with the beam specimen. From these base 3D mesh models, deformed 3D meshes are generated by applying analytical solutions for the deflection of cantilever beams under a point load, which is given by:

$$\begin{cases} y(x) = -\dfrac{P}{6EI}\left(3lx^2 - x^3 - 3lx_0^2 + x_0^3\right), & x \le x_0 \\ y(x) = -\dfrac{P}{6EI}\left(3lx^2 - x^3 - 3lx_0^2 + x_0^3 + 3(x - x_0)\left(x_0^2 - lx_0\right)\right), & x > x_0 \end{cases} \tag{14}$$

where l is the length of the beam, E is the Young's modulus, x_0 is the location that load is applied and I is the moment of inertia of the beam's cross-section. Point loads are applied at 20 uniformly spaced locations

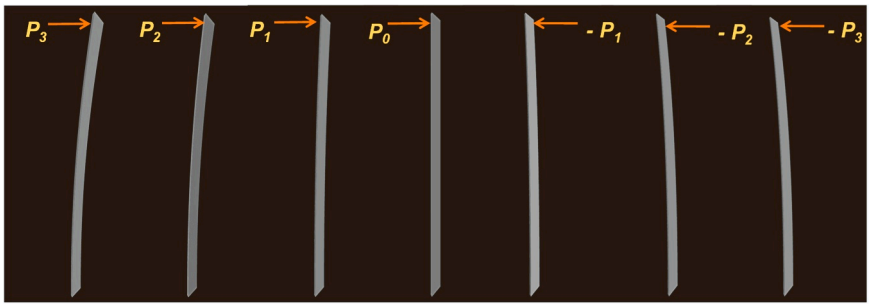


Fig. 8. Training data examples generated by 3DGEN. Illustrating meshes subjected to varying levels of force at the free end.

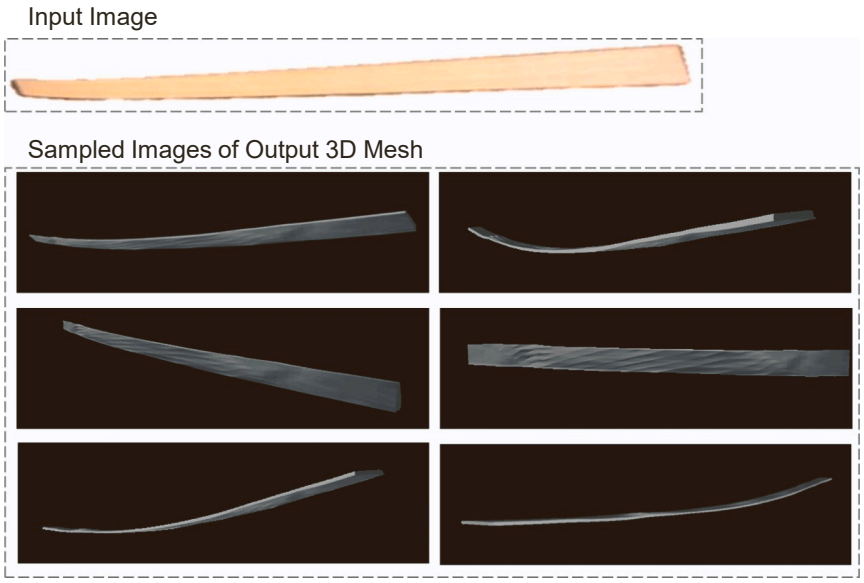


Fig. 9. Rendered images from various viewing angles of a 3D mesh reconstructed from a single image of the wooden beam.

along the beam structures' length. At each of these locations, 250 distinct loading intensities are employed. For every individual mesh, images are captured from 100 different viewpoints. Cumulatively, the dataset comprises 5000 deformed meshes and 500 K images. All 3D meshes undergo normalization using the min-max method [33]. Each generated 3D mesh comprises 6146 vertices, a measure of mesh granularity that ensures comprehensive full-field displacement measurements of simple beam structures. This vertex density is specifically selected based on the complexity of the structure being modelled and the precision required for accurate displacement analysis. The data generation was conducted using the Pyrender API and was completed about 10 h on a 12th Gen Intel(R) Core (TM) i7–1255 U CPU. Some examples of the training data are shown in Fig. 8.

3.3. Results of wood beam vibration testing

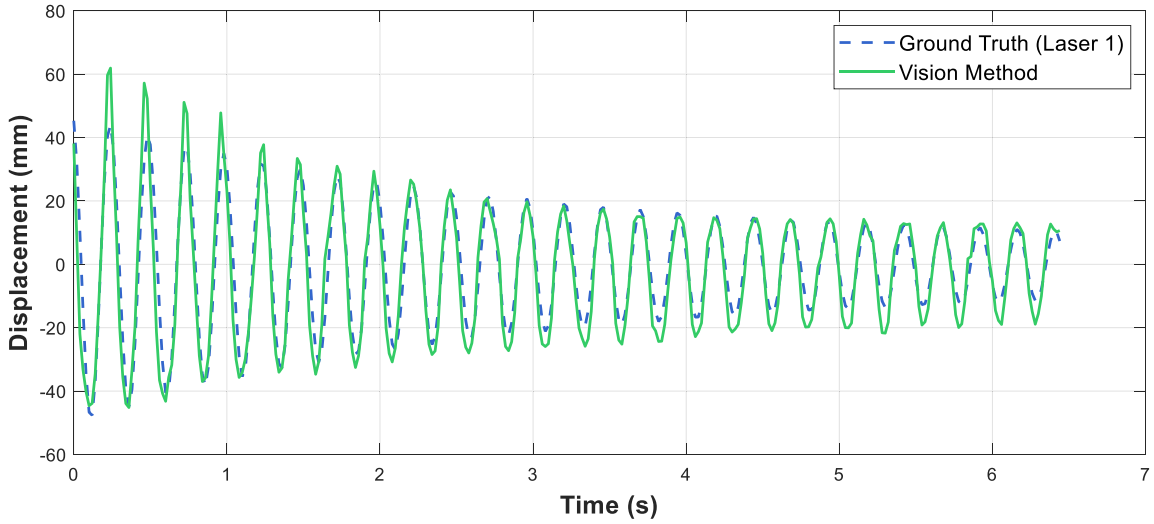
Fig. 9 presents a sample of a reconstructed 3D mesh of the wooden beam. At the top of the figure, a segmented input image is displayed, while. While certain detailed portions of the reconstructed mesh exhibit minor issues, the overall 3D form reconstructed is commendably accurate.

LDSs are used at two specific points on the actual beam structure, and corresponding points on the reconstructed mesh are identified and tracked. These selected vertices within the mesh are consistently monitored across all reconstructed meshes throughout the video. Displacement measurements are derived by calculating the changes in these vertex positions from one frame to another, ensuring that the measurements derived from the vision-based method closely align with the LDS ground truth. Fig. 10 compares the displacement histories measured by the proposed vision-based method with the ground truth displacements captured by LDS. The green lines signify the results from the proposed method, while the blue dashed lines depict the LDS recorded ground truth. Fig. 10(a) shows the displacement time history of a point in the area where Laser 1 is installed, while Fig. 10(b) shows that of a point from the Laser 2 installation area. It can be clearly observed that the measurements derived from the vision-based method closely align with the ground truth.

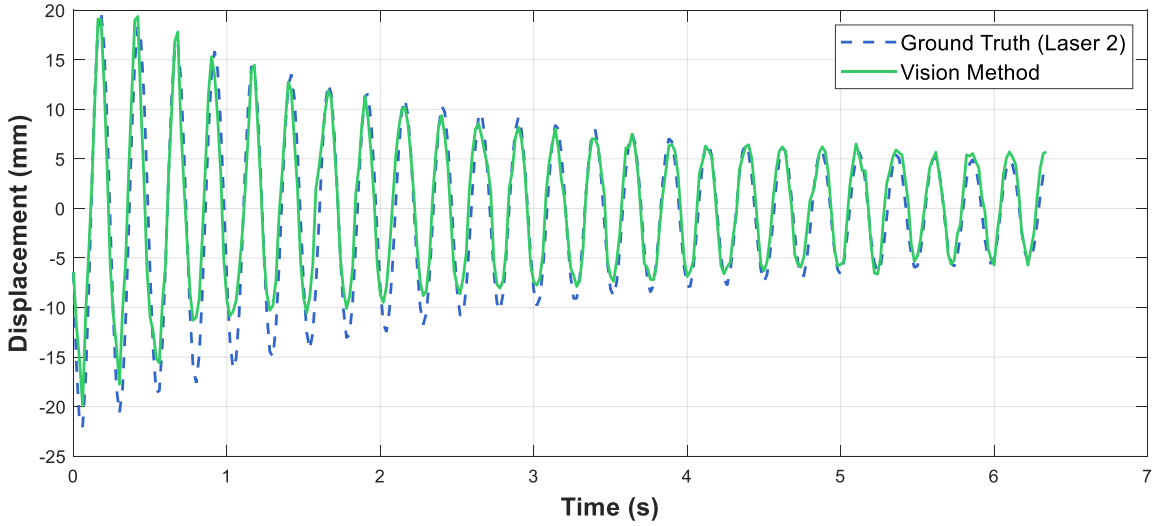
The Cross-Correlation Coefficient (CCF) and Mean Absolute Percentage Error (MAPE) between the measurement displacement time histories from the two methods are detailed in Table 1. CCF is defined as follows for displacement measurements a_i from sensors and b_i from proposed vision-based method:

$$CCF = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{a_i-\mu_a}{\sigma_a}\right)\left(\frac{b_i-\mu_b}{\sigma_b}\right), \tag{15}$$

where μ_a, μ_b are the means, and σ_a, σ_b are the standard deviations of signals a_i and b_i, respectively. The Cross-Correlation Function (CCF) produces values within the range of -1 to 1, where higher values signify increased similarity between the compared signals. MAPE values span from zero to infinity, with decreased values signifying greater accuracy. It assesses the relative errors in readings from vision-based methods in comparison to laser sensor readings.



(a)



(b)

Fig. 10. Time history of vibration measurement for the wooden beam at two locations: (a) Vision vs. Laser 1; (b) Vision vs. Laser 2.

Table 1
Displacement error analysis of wooden beam vibration test.

Location	CCF	MAPE (%)
Laser 1	0.9430	29.97
Laser 2	0.9513	28.98

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{a_i - b_i}{a_i}\right| \times 100\%. \tag{16}$$

It is evident from the table that the proposed measurement system delivers accurate measurements. The CCF exceeds 0.94, and the MAPE remains below 30 %.

The time history displacement data, measured by the proposed vision system and lasers, is processed using the Fast Fourier Transformation (FFT) to transition to the frequency domain. This assists in pinpointing the vibration frequencies of the structure. As illustrated in Fig. 11, the proposed vision system successfully identifies vibration frequencies, which align closely with those captured by the lasers.

3.4. Results of aluminium beam vibration testing

Fig. 12 displays the reconstructed 3D mesh of the aluminium beam. A notable observation from the visualization is that the thickness of the beam has not been accurately captured. The reconstructed 3D representation appears flattened. This discrepancy can be attributed to the thinness of the aluminium beam, which is just 3 mm, making it challenging to reconstruct the height accurately. However, it is observed that this discrepancy in reconstruction does not cause major issue later for the estimation of the full-body deformation.

Fig. 13 presents the time history of the vibration of the aluminium beam captured by the vision-based approach and the Laser sensors, in which 12(a) and 12(b) show the performance in the area of Laser 1 and Laser 2 respectively. It can be clearly seen that the readings from the vision-based approach demonstrate a strong correlation with the ground truth. An analysis of the CCF and MAPE between both data trajectories can be found in Table 2. Fig. 14 depicts the frequency domain displacement of the aluminium beam. The obtained fundamental vibration frequency, measured at 2.4390 Hz, is exactly the same by the

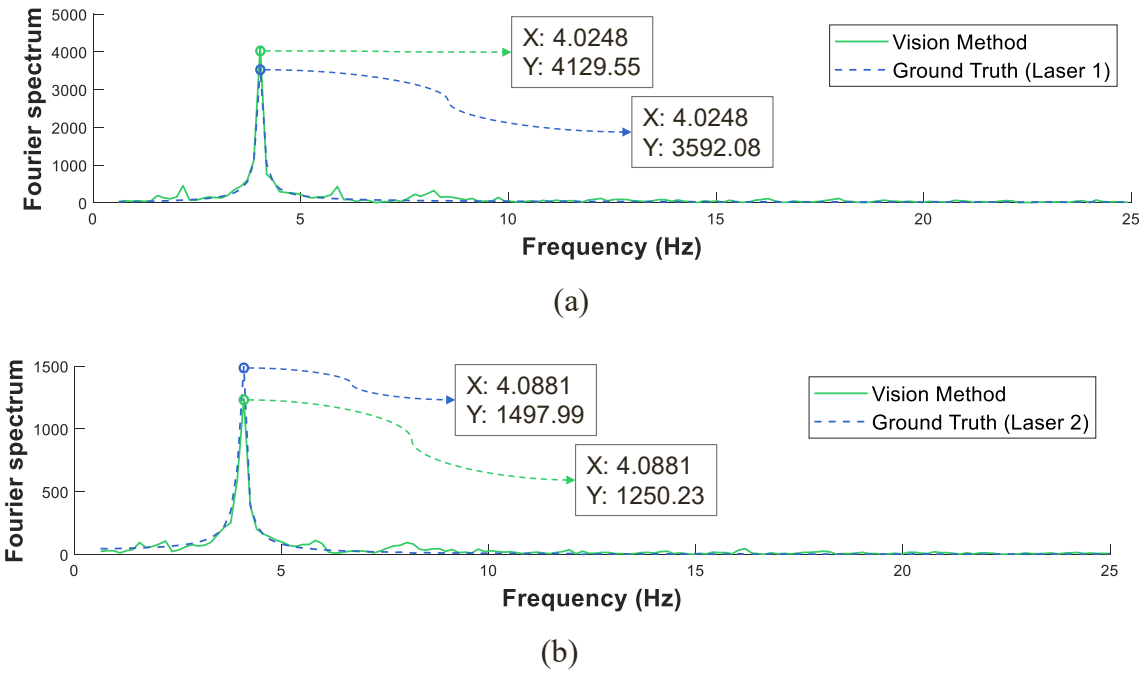


Fig. 11. FFT spectra of wooden beam time history vibration displacements measured by the proposed vision approach and displacement sensors: (a) Laser 1 vs. vision method; (b) Laser 2 vs. vision method.

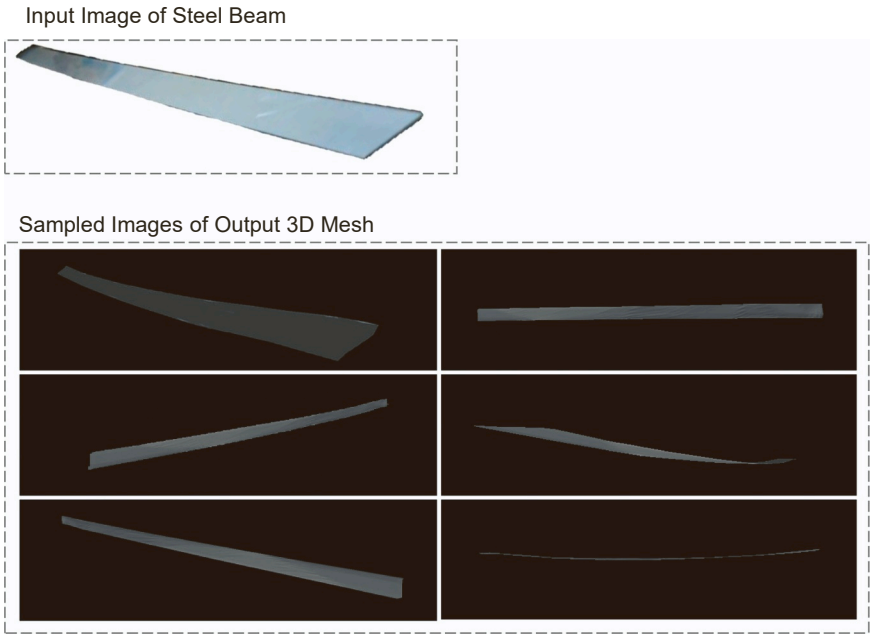
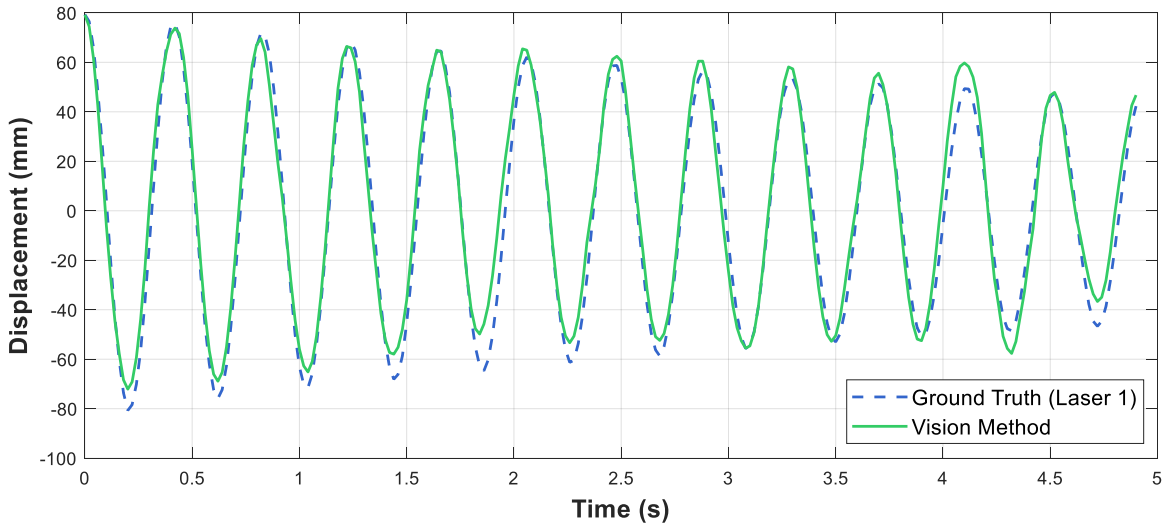


Fig. 12. Rendered images from various viewing angles of a 3D mesh reconstructed from a single image of the aluminium beam.

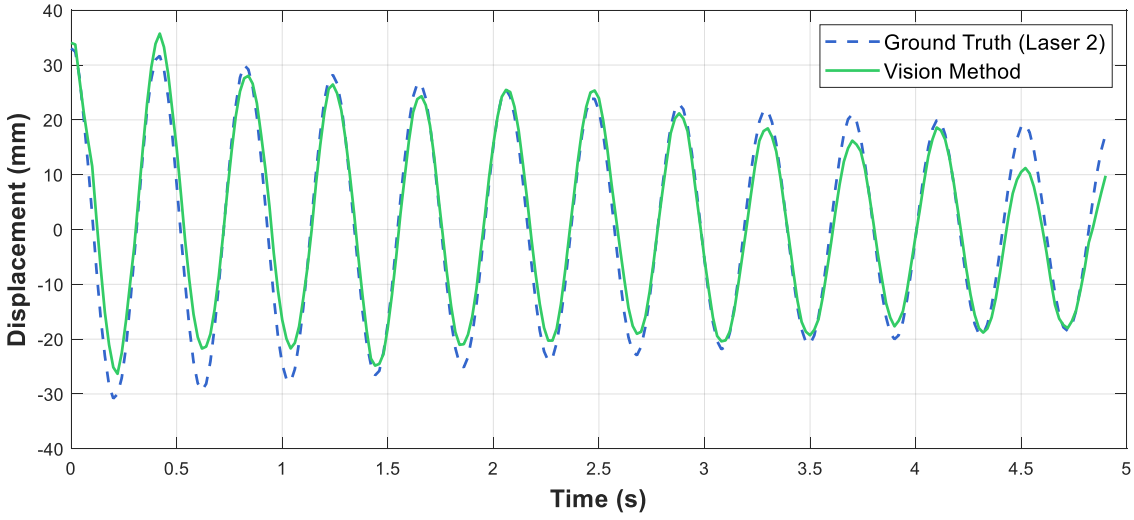
two methods.

4. Conclusion

In this paper, a target-free method for 3D displacement measurement using a monocular camera is introduced. This method is entirely target-free, eliminating the dependence on artificial markers or natural key points. A novel mesh deformation neural network is trained on synthetic data, allowing for the prediction of displacement at any point on the structure. Furthermore, an advanced video segmentation neural network is utilized to isolate structures from background, enhancing the robustness of the measurement system. The efficacy of this approach is validated through two experimental tests on beam structures. The measured displacement time histories closely align with the ground truth obtained by conventional lasers, with CCF exceeding 0.94 and MAPE staying below 30 %. The observed high MAPE can be attributed to a combination of factors related to equipment specifications and computational constraints. Notably, there is an inherent discrepancy in the data acquisition rates, with the camera operating at 155.83fps and the laser at 200 Hz. Additionally, due to the current computational limitations associated with the Tracking Anything [32] segmentation process, the camera's resolution is reduced, necessitating the down-sampling of videos to 50fps. Concurrently, the ground truth data from laser is adjusted to 50 Hz to align with the down-sampled videos.



(a)



(b)

Fig. 13. Displacement time histories of the aluminium beam: (a) Vision vs. Laser 1; (b) Vision vs. Laser 2.

Table 2
Error analysis of aluminium beam displacement measurements.

No.	CCF	MAPE (%)
Laser 1	0.9812	18.35
Laser 2	0.9810	23.43

This down-sampling process introduces synchronization challenges between the laser data and video, potentially leading to the observed larger percentage errors.

While the proposed method offers a significant step forward in displacement monitoring of civil structures, there are many ways it can be further improved. The following points encapsulate the key observations and suggest trajectories for future research that could potentially amplify the method's practicability. **1) Complex Structures:** While the current network performs admirably in measuring the responses of simple structures such as beams, it struggles with more intricate structures like transmission towers. The network can only approximate a general shape for those structures, falling short in precisely measuring their displacement. A future priority should be refining the network's capability to comprehend for vibration measurements of more complex structures accurately. **2) Vibration Video Constraints:** For the method to work effectively, the entire structure must be visible in the vibration video, with a few fixed reference points on the structure. This restriction potentially limits the method's applicability to small or medium-sized structures. Adapting the network to infer deformations from partially obscured images to large structures, would be a valuable advancement. **3) Measurement of Minor Displacements:** Another area for enhancement is the capability to measure minor displacements, which are often critical in damage detection analysis. Current limitations include the resolution of the camera system and the neural network's ability to discern subtle changes in the mesh deformation. Future improvements could involve integrating higher resolution imaging technologies and enhancing the neural network's sensitivity through advanced training techniques that focus on minor movements.

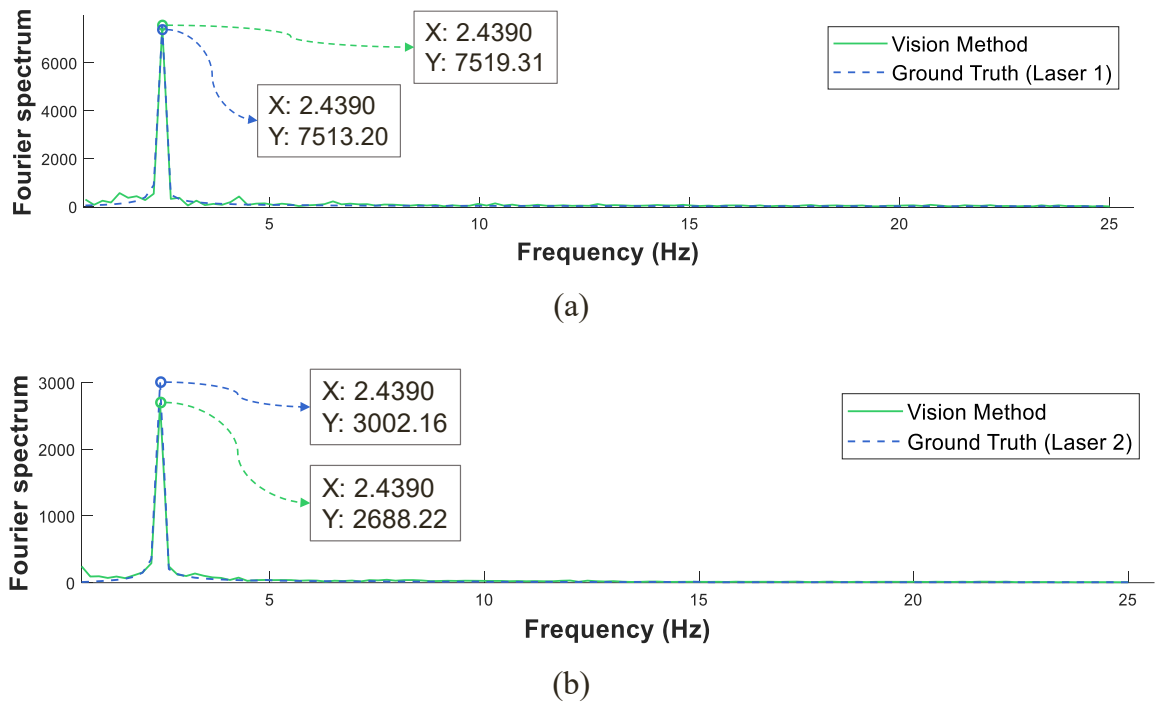


Fig. 14. FFT spectra of aluminium beam time history vibration displacements measured by the proposed vision approach and displacement lasers: (a) Laser 1 vs. vision method; (b) Laser 2 vs. vision method.

CRediT authorship contribution statement

Yanda Shao: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Ling Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Senjian An:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Hong Hao:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Jun Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Qilin Li:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

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

Data Availability

Data will be made available on request.

Acknowledgement

The support from Australia Research Council Discovery Project DP210103631: AI Assisted Probabilistic Structural Health Monitoring with Uncertain Data, is acknowledged.

Declaration of conflicting interests

The authors declare that there is no conflict of interest.

References

[1] Feng D, Feng MQ. Vision-based multipoint displacement measurement for structural health monitoring. Struct Control Health Monit 2016;vol. 23(5):876–90.
[2] Hao H, Bi K, Chen W, Pham TM, Li J. Towards next generation design of sustainable, durable, multi-hazard resistant, resilient, and smart civil engineering structures. Eng Struct 2023;vol. 277:115477.
[3] Feng D, Feng MQ. Computer vision for SHM of civil infrastructure: from dynamic response measurement to damage detection–A review. Eng Struct 2018;vol. 156: 105–17.
[4] Dong C-Z, Celik O, Catbas FN, O'Brien EJ, Taylor S. Structural displacement monitoring using deep learning-based full field optical flow methods. Struct Infrastruct Eng 2020;vol. 16(1):51–71.
[5] Kuddus MA, Li J, Hao H, Li C, Bi K. Target-free vision-based technique for vibration measurements of structures subjected to out-of-plane movements. Eng Struct 2019; vol. 190:210–22.
[6] Ribeiro D, Calçada R, Ferreira J, Martins T. Non-contact measurement of the dynamic displacement of railway bridges using an advanced video-based system. Eng Struct 2014;vol. 75:164–80.
[7] Giri P, Kharkovsky S. Detection of surface crack in concrete using measurement technique with laser displacement sensor. IEEE Trans Instrum Meas 2016;vol. 65 (8):1951–3.
[8] Yi TH, Li HN, Gu M. Recent research and applications of GPS-based monitoring technology for high-rise structures. Struct Control Health Monit 2013;vol. 20(5): 649–70.
[9] Cinque D, Saccone M, Capua R, Spina D, Falcolini C, Gabriele S. Experimental validation of a high precision GNSS system for monitoring of civil infrastructures. Sustainability 2022;vol. 14(17):10984.
[10] Manzini N, et al. Performance analysis of low-cost GNSS stations for structural health monitoring of civil engineering structures. Struct Infrastruct Eng 2022;vol. 18(5):595–611.
[11] Shao Y, Li L, Li J, An S, Hao H. Computer vision based target-free 3D vibration displacement measurement of structures. Eng Struct 2021;vol. 246:113040.

[12] Busca G, Cigada A, Mazzoleni P, Zappa E. Vibration monitoring of multiple bridge points by means of a unique vision-based measuring system. Exp Mech 2014;vol. 54:255–71.
[13] Miao Y, Kong Y, Nam H, Lee S, Park G. Phase-based vibration imaging for structural dynamics applications: Marker-free full-field displacement measurements with confidence measures. Mech Syst Signal Process 2023;vol. 198: 110418.
[14] Tan D, Li J, Hao H, Nie Z. Target-free vision-based approach for modal identification of a simply-supported bridge. Eng Struct 2023;vol. 279:115586.
[15] Wang Y, Hu W, Teng J, Xia Y. Phase-based motion estimation in complex environments using the illumination-invariant log-Gabor filter. Mech Syst Signal Process 2023;vol. 186:109847.
[16] Lydon D, Lydon M, Taylor S, Del Rincon JM, Hester D, Brownjohn J. Development and field testing of a vision-based displacement system using a low cost wireless action camera. Mech Syst Signal Process 2019;vol. 121:343–58.
[17] Li M, Wang S, Liu T, Liu X, Liu C. Rotating box multi-objective visual tracking algorithm for vibration displacement measurement of large-span flexible bridges. Mech Syst Signal Process 2023;vol. 200:110595.
[18] Ma Z, Choi J, Sohn H. Three-dimensional structural displacement estimation by fusing monocular camera and accelerometer using adaptive multi-rate Kalman filter. Eng Struct 2023;vol. 292:116535.
[19] Park S, Park HS, Kim JH, Adeli H. 3D displacement measurement model for health monitoring of structures using a motion capture system. Measurement 2015;vol. 59:352–62.
[20] Shao Y, Li L, Li J, An S, Hao H. Target-free 3D tiny structural vibration measurement based on deep learning and motion magnification. J Sound Vib 2022; vol. 538:117244.
[21] Wang C, Xiao T, Gong Z, Yang S, Zhang D, Deng F. Wireless Binocular Stereovision Measurement System Based on Improved Coarse-to-Fine Matching Algorithm. Struct Control Health Monit 2023;vol. 2023.
[22] Lee J, Lee K-C, Jeong S, Lee Y-J, Sim S-H. Long-term displacement measurement of full-scale bridges using camera ego-motion compensation. Mech Syst Signal Process 2020;vol. 140:106651.
[23] Lee S, Kim H, Sim S-H. Equation Chapter 1 Section 1 nontarget-based displacement measurement using LiDAR and camera. Autom Constr 2022;vol. 142:104493.
[24] Javed A, Park J, Lee C, Lee H, Kim B, Han Y. Edge-based 3D vibration measurement of rotating cylinder-shaped structure through epipolar line-based corresponding point extraction between two camera images. Mech Syst Signal Process 2023;vol. 187:109981.
[25] Chang C, Xiao X. Three-dimensional structural translation and rotation measurement using monocular videogrammetry. J Eng Mech 2010;vol. 136(7): 840–8.
[26] Shao Y, Li L, Li J, Li Q, An S, Hao H. Monocular vision based 3D vibration displacement measurement for civil engineering structures. Eng Struct 2023;vol. 293:116661.
[27] Sun C, Gu D, Lu X. Three-dimensional structural displacement measurement using monocular vision and deep learning based pose estimation. Mech Syst Signal Process 2023;vol. 190:110141.
[28] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;vol. 60:91–110.
[29] Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF). Comput Vis Image Underst 2008;vol. 110(3):346–59.
[30] P.F. Alcantarilla A. Bartoli A.J. Davison KAZE features Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12 2012 Springer, 214 227.
[31] D. DeTone, T. Malisiewicz, and A. Rabinovich, Superpoint: Self-supervised interest point detection and description, in Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 224–236.
[32] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, and F. Zheng, Track anything: Segment anything meets videos, *arXiv preprint arXiv:2304.11968,* 2023.
[33] Shao Y, Li L, Li J, Li Q, An S, Hao H. 3DGEN: a framework for generating custom-made synthetic 3D datasets for civil structure health monitoring. Structural Health Monitoring 2024:14759217241265540.
[34] K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[35] A.F. Agarap, Deep learning using rectified linear units (relu), *arXiv preprint arXiv: 1803.08375*, 2018.
[36] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, Pixel2mesh: Generating 3d mesh models from single rgb images, in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 52–67.
[37] D.P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980,* 2014.
[38] Besl PJ, McKay ND. Method for registration of 3-D shapes. Sensor fusion IV: control paradigms and data structures, vol. 1611. Spie,; 1992. p. 586–606.
[39] A. Kirillov et al., Segment anything, arXiv preprint arXiv:2304.02643, 2023.
[40] Cheng HK, Schwing AG. Xmem: long-term video object segmentation with an atkinson-shiffrin memory model. European Conference on Computer Vision. Springer; 2022. p. 640–58.