Automation in Construction 165 (2024) 105507

Contents lists available at ScienceDirect

Automation in Construction

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

Out-of-plane full-field vibration displacement measurement with monocular computer vision

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:
3D displacement measurement
Structural health monitoring
Civil engineering
Computer vision
Deep learning
Monocular vision

ABSTRACT

Vibration displacement of civil structures is crucial information for structural health monitoring (SHM). The challenges and costs associated with traditional physical sensors make displacement measurement not always straightforward owing to difficulties such as inaccessibility. While recent computer vision based methods for displacement measurements offer simplicity, unfortunately they lag in terms of accuracy and robustness. This paper introduces a monocular camera system designed to measure out-of-plane vibration displacement. Compared to existing monocular-camera based methods, the proposed monocular vision-based measurement technique significantly enhances accuracy and robustness. This boost can be attributed to the generation of a vast and precise dataset and augmented by employing advanced techniques for object segmentation and background elimination. Experimental tests are conducted in the laboratory to investigate the feasibility of the proposed system. The results demonstrate that the proposed monocular 3D displacement system can produce highly accurate full-field out-of-plane displacement measurement.

1. Introduction

In recent decades, the application of computer vision technologies in civil structural displacement measurement has attracted substantial interest, owing to their notable advantages over traditional methods like displacement sensors. Vision-based measurement methods offer a non-contact approach that significantly surpasses traditional contact-based techniques in efficiency and applicability. These methods eliminate the need for physical interaction with the structure, thereby reducing the risk of causing any damage during sensor installations and enabling the monitoring of otherwise inaccessible or sensitive areas without direct contact. Vision-based displacement measurement methods offer significantly higher accuracy and resolution compared to Global Navigation Satellite System (GNSS) or radar-based techniques. Furthermore, they can provide rich visual information that allows for a more comprehensive analysis of the object's surface characteristics and deformations. The cost-effectiveness and ease of deployment associated with camera-based monitoring present a scalable solution that can significantly reduce the logistical complexity and overall expense of SHM operations. A single camera setup can simultaneously monitor multiple points on a structure, offering a multi-point measurement capability that would otherwise require an extensive network of individual sensors. Additionally, the capacity for remote monitoring inherent to vision-based methods greatly enhances the flexibility and reach of SHM practices, particularly for structures located in hazardous or difficult-to-access areas. The integration of vision-based data with advanced image processing and machine learning algorithms further extends the capabilities of SHM, enabling automated anomaly detection, pattern recognition, and predictive maintenance strategies. Various methods have been developed in the field of computer vision-based displacement measurement. Among these, in-plane displacement measurement methods have been explored, developing techniques to accurately capture movements within a single plane [1–5]. Others have extended this research to achieve out-of-plane displacement measurement, addressing the complexities of movements outside the primary plane of observation [6–10]. Some studies have focused on target-free methods, offering flexibility in applications without the need for specific markers [11–14]. An area of great research interests has been the use of monocular cameras for out-of-plane displacement measurement, leveraging a single camera to capture three-dimensional information

* 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.autcon.2024.105507
Received 7 November 2023; Received in revised form 1 April 2024; Accepted 31 May 2024
Available online 5 June 2024

[15–17]. Additionally, some studies employed Unmanned Aerial Vehicles (UAVs) for displacement measurement, providing unique perspectives and access to challenging locations [18,19]. A major benefit of these cutting-edge techniques lies in their growing accessibility, cost-effectiveness, and features that enhance their practicality across various applications. Unlike traditional displacement sensors, computer vision provides the opportunity for non-contact, straightforward setting-up and cost-effective measurement. This attribute is particularly valuable in monitoring structures that are delicate or sensitive to disturbance. Moreover, the ability of computer vision for remote sensing permits measurements to be taken from a distance without the need of direct access to the structure. This capability is ideal for large-scale, remote and inaccessible areas, setting it apart as a convenient tool for modern civil engineering practices.

In the context of vision-based displacement measurement, it is essential to distinguish between in-plane and out-of-plane displacements. In-plane displacement refers to the displacement that predominantly occurs within the plane of the image, primarily along the X and Y directions (horizontally and vertically within the image plane) of the camera coordinate system. On the other hand, out-of-plane displacement pertains to displacement occurring in the depth direction, also known as the *Z*-direction of the camera coordinate system. When the structure in the world coordinate system is moving in a one/two-dimensional (1D or 2D) plane, it is possible to conduct the measurement as an in-plane displacement measurement without losing any information. This is because the problem becomes a 2D to 2D mapping problem, allowing for a straightforward representation of the displacement within the captured images/frames. In instances of in-plane displacement measurement, a crucial step is strategically aligning the camera plane to directly face the plane of displacement. This alignment minimizes the depth-related complexities and aids in accurately capturing the in-plane displacement. Complementing this approach, another crucial component involves correcting perspective distortions and scaling the captured image displacements to real-world units (e.g., mm) [3]. This is achieved through techniques such as homographic transformations or precise measurements of the camera's position and orientation. These combined strategies aid in transforming the captured displacements into a coherent representation of the real-world movements.

In certain scenarios, displacement in the depth direction (out-of-plane displacement) could occur and become unavoidable. This is a frequently encountered scenario in real-world, where practical constraints such as the case when the structure's movement in the world coordinate system is inherently three-dimensional (3D). Furthermore, even when the displacement in the world coordinate system is restricted to 1D or 2D, practical constraints such as accessibility to the site and visibility of the structure can still hinder optimal camera placement. Despite careful alignment and the implementation of correction techniques (e.g., homographic transformations), the effect of the perspective distortion caused by depth direction movement may be impossible to eliminate. These situations necessitate the methods that carefully consider out-of-plane displacement, emphasizing the importance of comprehensive strategies to capture the full spectrum of civil structures' movement. Capturing the out-of-plane displacement poses significant challenges, as the imaging process inherently loses depth information. Accurately measuring out-of-plane displacement generally needs additional information to recover the lost depth information. Advanced techniques are leveraged to recover the depth information and reconstruct the 3D scene from the 2D images. Multiple camera-based displacement methods [7,9,10,12,14] involved the use of multiple cameras placed at different viewpoints to capture the scene. These cameras provide different perspectives and enable the estimation of 3D displacement of objects or features. It is based on the principle that when the same scene or object is observed from different viewpoints, the displacement in the captured images can be used to compute the depth information or the third-dimensional (out-of-plane) displacement. Typically, at least two cameras are required, placed at known positions relative to one another. More cameras can be added to improve accuracy or capture more viewpoints, but this adds to the complexity. Before measurements can be made, the camera systems must be calibrated. This involves determining the relative positions and orientations of the cameras, the focal lengths, and any lens distortions. Once calibrated, the 3D position of a point in space can be determined by finding the intersection of the rays coming from each camera to that point. This process, known as triangulation [20], allows for the calculation of depth information. Pan and Yang [9] introduced a deep learning based computer vision-based framework for measuring out-of-plane displacements of steel plate structures. The framework leverages multi-view vision algorithms and deep learning to create a comprehensive 3D point cloud representation of the structures and their surrounding environment. Park et al. [10] proposed a motion capture system (MCS) as a versatile tool that can precisely determine marker movements in any direction. Unlike traditional 1D or 2D displacement sensors, MCS can overcome frequency sampling limitations often encountered with terrestrial laser scanning (TLS) and global positioning systems (GPS). Utilizing multiple cameras, MCS measures the 2D coordinates of various markers, translating them into 3D coordinates. Shao et al. [14] proposed a method for 3D vibration displacement measurement of civil engineering structures, using a binocular vision system. The proposed vision-based method leverages deep learning algorithms for key point detection and matching, enabling target-free measurement. The work is later refined [14] to optimize its capability for micro displacement detections.

Multi-view camera systems, though adept at providing depth information for out-of-plane displacement measurements, are not without their challenges. Notably, they grapple with issues of occlusions, where features discernible in one camera might be obscured in another, complicating the measurement process. Rigorous and precise calibration is imperative, and even minor deviations can lead to substantial inaccuracies. Maintaining perfect synchronization among all cameras, especially in dynamic settings, is another challenge, slight misalignments may cause data disparities. The inherent complexity and cost of establishing such systems are further compounded by the need for enhanced computational capacity to process data from multiple sources in real-time. Spatial constraints can further complicate the setup, especially in constricted environments or locations with obstructions. Constricted environments inherently limit the available space to position cameras, making optimal placement a meticulous endeavor. Overcoming these challenges is often a prerequisite for successful deployment in real-world applications.

Monocular vision-based out-of-plane displacement measurement technologies are seeking to mitigate the challenges associated with multi-camera-based systems. Monocular vision based 3D displacement measurement method [15–17] is a significant advancement in computer vision that enables the inference of depth information from a single 2D image. It has emerged as a powerful tool in the fields of civil engineering and structural analysis. Monocular out-of-plane displacement measurement techniques are primarily rooted in two distinct approaches. Firstly, there are methods centred around hand-crafted depth cues, commonly termed as "shape-from-x" techniques. These methods attempt to infer 3D structure from 2D images based on certain visual cues, such as shading [21] and focus [22]. Examples include "shape-from-shading", where the variations in image brightness hint at the object's depth or "shape-from-focus", using the degree of sharpness across images to determine depth. While these techniques have been foundational in computer vision, they come with certain limitations. Specifically, "shape-from-x" methods often rely heavily on ideal conditions. For instance, accurate depth perception using shading requires consistent lighting, and any deviation can lead to errors. On the other hand, deep learning-based methods, a more recent innovation, harness the power of neural networks. This category can be further broken down into techniques like depth estimation [17], which directly predicts depth maps from images, and 3D object detection methods [16],

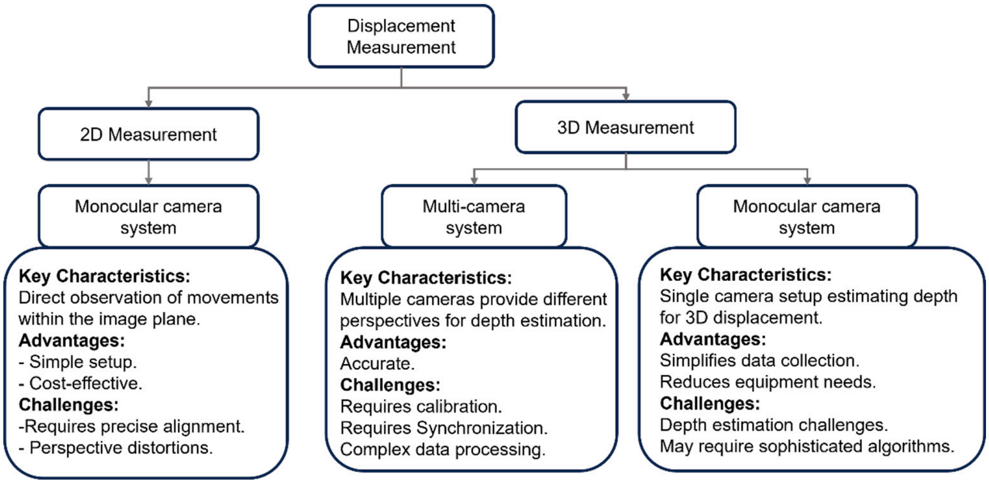


Fig. 1. A taxonomy diagram of computer vision technologies for structure displacement.

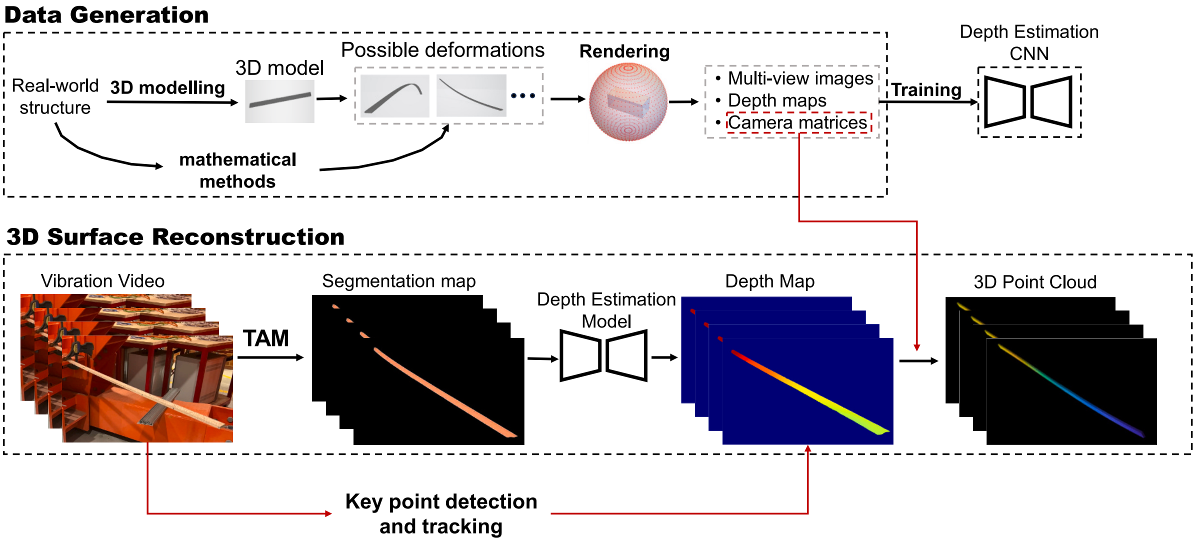


Fig. 2. The overall pipeline of the proposed monocular vision based out-of-plane displacement measurement method.

focusing on recognizing and spatially locating 3D objects in 2D images.

The adaptability and learning capability of these methods have made them increasingly popular in modern applications. Sun et al. [16] proposed a 3D structural displacement measurement method using monocular vision and deep learning-based pose estimation, which has demonstrated a commendable level of measurement accuracy. The method utilizes virtual rendering to synthesize a training set based on 3D models of target objects, trains a deep learning model to estimate target object poses, and measures 3D translations of structures based on the original, and destination poses or key point matching. Presently, this method predominantly focuses on locating rigid body structures. This poses a significant limitation when it comes to civil structures, which inherently exhibit flexible characteristics. Unlike idealized rigid bodies, civil infrastructures like bridges have elements that respond differently under various conditions. While the columns of a bridge, anchored firmly to the ground, experience relatively small motion under regular circumstances, the beam components however could exhibit relatively larger motions under regular loading conditions such as traffic and wind. Under extreme event of significant disturbances such as a major earthquake, both columns and beams may experience large and complex vibrations. It is crucial to note that once a structure incurs extensive damage, measuring its vibrations may help to quickly assess the structural conditions. Another prominent approach in this domain is the monocular depth estimation method. Shao et al. [17] pioneered a system that employs a singular camera for 3D vibration displacement measurements. Eschewing the conventional multi-view geometry, this system harnesses the power of deep neural networks to infer depth from monocular images. However, its efficacy, in terms of accuracy, remains an area to be improved. A significant challenge with this depth-estimation method is the stark relative error, often escalating to 50% or even higher. The crux of this issue lies in the absence of training data specific to civil structures and the (lack of) precision of the depth maps used for training. Conceptually, the method holds promise for non-rigid body displacement measurements. In the common datasets used for training depth estimation networks, a limitation is observed. Most of these datasets are primarily focused on static indoor items, so the ability to understand deformations is not effectively imparted. As a result, the trained network becomes proficient in identifying distances between various objects, but the nuances of how a single object might deform are often overlooked. This deficiency poses challenges when the network is applied to flexible structures buildings or bridges. While some datasets including outdoor scenes are available, they are frequently compromised in quality, often being generated through dual-camera systems. Consequently, accurate measurements are hard to achieve when networks are trained on these suboptimal datasets. A flowchart summarizing the technologies for computer vision-based structural

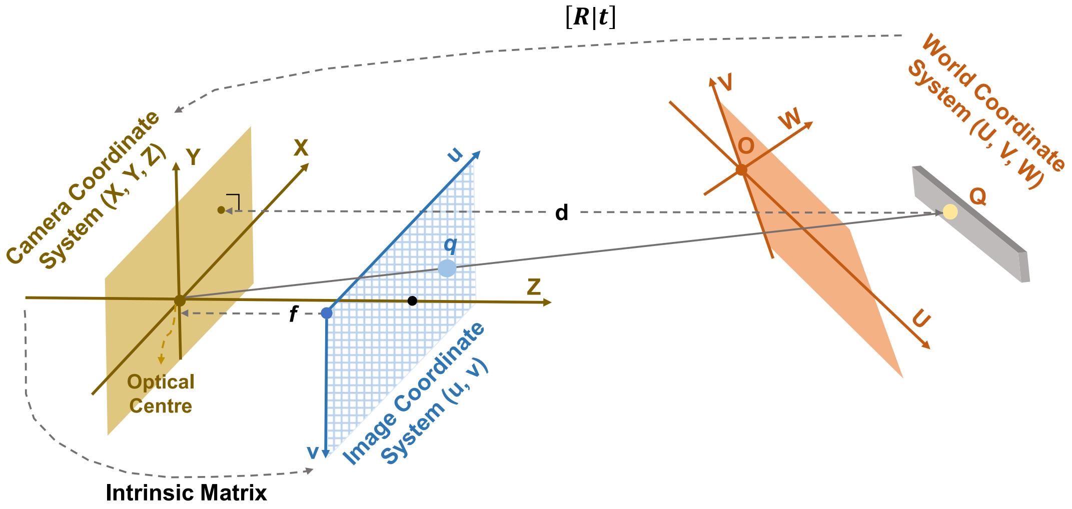

[R|t]
World Coordinate System (U, V, W)
Camera Coordinate System (X, Y, Z)
Y
X
Z
u
v
q
d
f
O
Q
V
W
U
Optical Centre
Image Coordinate System (u, v)
Intrinsic Matrix

Fig. 3. Monocular pinhole camera model.

displacement measurement is presented in Fig. 1 to facilitate a clearer understanding of the methodologies discussed.

The system presented in this paper seeks to improve the precision of non-rigid body out-of-plane vibration displacement measurements for civil structures. The overview of this measurement system is shown in Fig. 2. The proposed system can be classified into depth estimation based out-of-plane displacement measurement. To enhance measurement accuracy, a detailed dataset tailored to specific structural needs is generated. This dataset is generated using a 3D data generation method designed for rapid, large-scale 3D civil structure data production. Given minimal prior knowledge about the structure, such as boundary conditions and dimensions, the method can predict a plethora of potential deformations. Utilizing numerical tools and analytic solutions, this vast dataset, representing a multitude of possible structural deformations, can be generated in as little as a week (millions of images and depth maps). By training a depth estimation neural network on this extensive dataset, a "depth estimation expert" primed for future is created, including unforeseen displacement scenarios of the interested structure. The estimated depth, when combined with the intrinsic parameters from the 3D model renderings, paves the way for accurate 3D point reconstructions for each pixel. For every pixel within a video frame, there corresponds a reconstructed 3D point in a spatial dimension. To determine its displacement, the pixel's 3D location must be identified across all frames. This is usually achieved using a key point-based displacement measurement method [6]. Initially, key points are identified in the inaugural frame, and subsequently, these points are tracked across subsequent frames. By aligning the location of these key points on the frame with their respective positions on the depth map, the 3D trajectory of each key point throughout the video sequence can be pinpointed. Generating intricate backgrounds for 3D data can be computationally taxing. Previous research [23] indicates that crafting a pair of images having accurate background depth could take up to 3.5 h using a single NVidia GPU. Moreover, as highlighted earlier, some civil structures typically exhibit certain flexibility. This means that during vibrations, structures can manifest a multitude of mode shapes. The sheer diversity of these mode shapes provides ample data to train a network in deducing the depth of various points on the structure. Such a varied dataset enables networks to adeptly learn and recognize depth features by discerning patterns from these myriad shapes. Essentially, the network can derive depth directly from the structure of interest, eliminating the need to infer depth based on its relationship with background objects. Recognizing this bottleneck, the proposed measurement system employs advanced large vision models-Segment Anything (SAM) and Track Anything (TAM)-to efficiently extract the structure from individual frames, discarding the background. This strategic omission enables the network to be focus on structures.

The rest of paper is organized as follows: Section 2 details on the methodologies employed in the proposed vision-based displacement measurement system. This includes the principles of the pinhole camera model, the process of synthetic data generation, and the implementation of the depth estimation deep learning model. In Section 3, the effectiveness of the proposed system is assessed through two experimental vibration tests, which include comprehensive experimental validations, performance comparisons, and in-depth discussions. Finally, Section 4 concludes the paper, summarizing key findings and outlining directions for future work.

2. Methodology

2.1. Monocular vision system

A monocular camera system uses a single camera to capture images, providing a singular viewpoint similar to the human eye. This results in two-dimensional images derived from a three-dimensional scene, as shown in Fig. 3.

Central to understanding this system is the pinhole camera model, a mathematical framework detailing the geometric relationship between a 3D point (U,V,W) in the world coordinate system and its 2D projection (u,v) in the image coordinate system. The world coordinate system defines points in a global reference frame, while the camera coordinate system represents these points relative to the camera's position and orientation. In this idealized model, light traverses through a singular point, the "pinhole", projecting onto an imaging plane. The transformation from 3D world coordinates to 2D image coordinates can be described through the camera's projection matrix P, which is a combination of the camera's intrinsic matrix A and its extrinsic parameters (rotation matrix R and translation vector T):

$$P = A[R|T] \tag{1}$$

where,

$$A = \begin{bmatrix} \alpha & \gamma & u_0 \\ 0 & \beta & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

The intrinsic matrix A contains the camera's focal lengths along the $x(\alpha)$ and $y(\beta)$ axes, the principal point coordinates (u_0, v_0), and the skew factor γ. The extrinsic parameters are denoted by:

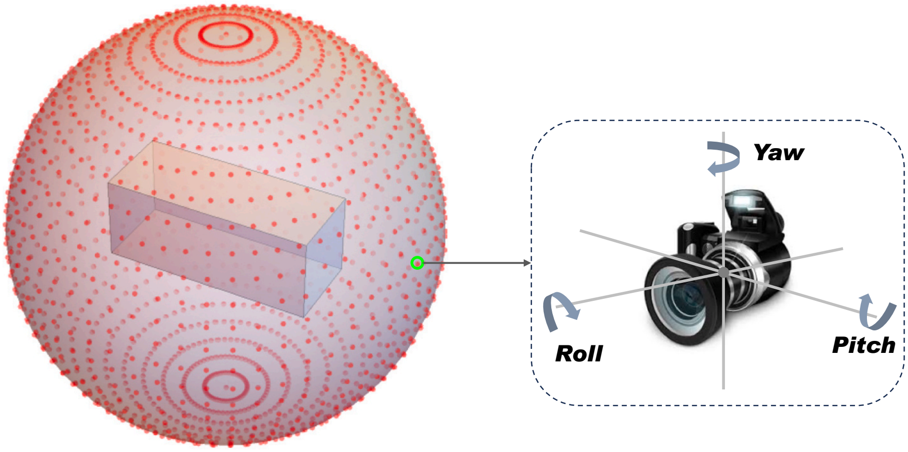


Fig. 4. The generation of camera positions by modifying a look-at matrix. The possible camera locations are represented by the red points. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

$$[R|T] = \begin{bmatrix} r_{11} & r_{21} & r_{31} & t_1 \\ r_{12} & r_{22} & r_{32} | & t_2 \\ r_{13} & r_{23} & r_{33} & t_3 \end{bmatrix} \tag{3}$$

Thus, the 2D homogeneous image coordinates $S[u \quad v \quad 1]^T$ can be obtained by: $S[u \quad v \quad 1]^T = A[R|T][U \quad V \quad W \quad 1]^T$. The scale factor S ensures the homogeneity of the resulting 2D coordinates.

One intrinsic limitation of this model, and monocular systems in general, is the loss of depth information during imaging. As the 3D scene gets mapped onto a 2D plane, the depth or the third dimension, which signifies the distance of objects from the camera, is not directly captured, since all points along a specific line of sight get mapped to a singular point on the imaging plane. However, if depth is available up to a scale and matches the camera parameters of a reference dataset, it becomes possible to reconstruct the 3D scene from 2D monocular images. Given depth information d (even up to a scale), the 3D coordinates (X, Y, Z) of a point in the camera coordinate system can be recovered from its 2D image coordinates (x, y) using

$$\begin{cases} X = \dfrac{x - u_0}{f} d \\ Y = \dfrac{y - v_0}{f} d, \\ \quad Z = d \end{cases} \tag{4}$$

Given the inherent limitations of monocular vision, when depth values for each pixel are acquired, even up to a scale, they can provide invaluable 3D information. To address the depth dimension lost during 2D image projection, depth maps are employed. In these maps, each pixel's value signifies the distance between the imaging plane and the corresponding point in the scene. By utilizing the depth information and camera's intrinsic matrix, each pixel in an image can be back-projected into the 3D space, resulting in an undistorted point cloud.

2.2. Data generation

Datasets play a foundational role in training neural networks, acting as the bedrock upon which models build their understanding and competence. For depth estimation networks, in particular, the quality and diversity of the depth maps in datasets can significantly influence the model's accuracy and generalization capabilities. Creating an exhaustive dataset that encompasses the vast majority of civil structures for training a generalized model for civil structure displacement measurement is an impractical endeavor. The sheer number of structures worldwide, each possessing its distinct features and traits, makes this task daunting. Moreover, when these structures undergo vibrations, the resulting mode shapes further diversify and compound the complexity. Given this immense variability, capturing a comprehensive snapshot of every conceivable displacement scenario becomes an insurmountable challenge. Crafting a bespoke and accurate image-depth dataset for each individual structure is a more pragmatic and effective approach. For instance, when designing a displacement measurement system for a bridge, instead of attempting to generalize from a vast dataset of various structures, a more focused approach would involve dedicating time to meticulously construct a 3D model of that specific bridge. Once this 3D model is in place, numerical modeling techniques, such as the Finite Element Method (FEM), can be employed to simulate potential vibrations the structure might encounter. This allows for a predictive analysis of how the structure will behave under various conditions. With the detailed 3D model and its deformations established, a precise image-depth dataset can be rendered, tailored specifically for that bridge. Training a neural network on this specialized dataset ensures that the resulting model is finely tuned to the nuances of that structure, maximizing accuracy and predictive capability. This approach not only streamlines the training process but also boosts the reliability and robustness of the measurement system when deployed in real-world scenarios.

The 3D dataset generation method, known as 3DGEN [24] , stands as a pivotal tool in crafting the essential training data. This dataset encompasses images and their corresponding depth maps. Firstly, an initial 3D model of the particular structure can be crafted based on design schematics or by directly measuring the dimensions of the structure. The higher the accuracy of this initial 3D model, the better the subsequent results and analyses. Central to this representation is the 3D mesh, capturing the intricate details of the interesting structure. To facilitate the generation of such a comprehensive representation, the Trimesh API [25] is used. This API is renowned for its capability to manage and process intricate 3D models. For modeling potential mode shapes of structures under various impacts, numerical techniques, such as analytic functions or the Finite Element Method (FEM), can be employed. Through these methods, the initial 3D mesh model undergoes deformation, resulting in a representation akin to structures subjected to external stimulation. The selection of numerical methods varies according to the specific structural type. For instance, in Section 3, analytic functions are utilized to craft a myriad of model shapes for cantilevers.

Utilizing the diverse 3D meshes of the structures under vibration that are created, the Blender rendering engine [26] is employed to generate both images and depth maps from synthetic cameras. The camera positions are randomly determined from points on a sphere that encircled the structure, as depicted in Fig. 4. This process resulted in the synthesis of multiple camera perspectives, showcasing the structures from a variety of angles. Correspondingly, for every camera perspective, the

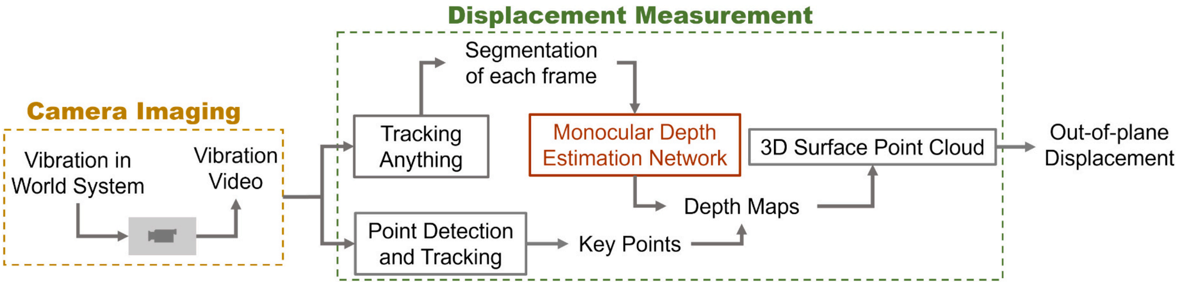

Displacement Measurement
Camera Imaging
Vibration in World System
Vibration Video
Segmentation of each frame
Tracking Anything
Monocular Depth Estimation Network
3D Surface Point Cloud
Out-of-plane Displacement
Depth Maps
Point Detection and Tracking
Key Points

Fig. 5. Overview of the proposed out-of-plane displacement measurement.

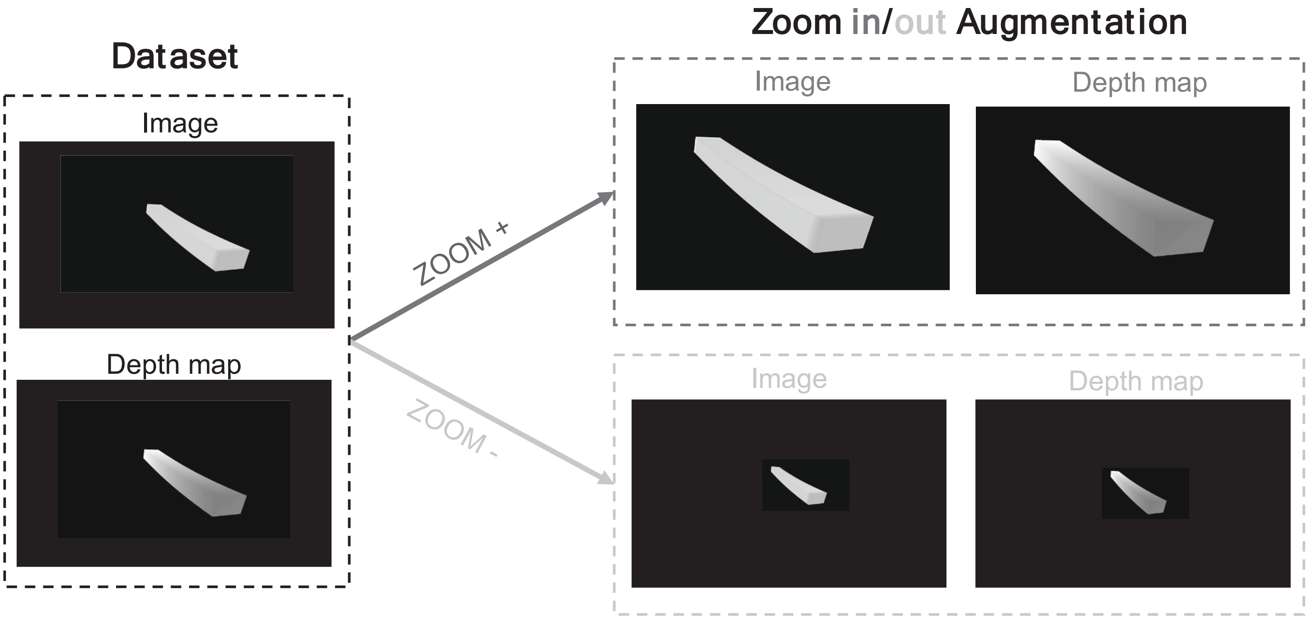


Fig. 6. A visualization of the zoom in/out augmentation. The pixel values of the depth maps are not changed by this operation.

intrinsic and extrinsic matrices of the camera are computed and retained. For simulated camera movement around the sphere, look-at matrices are generated by modifying the camera's roll, pitch, and yaw angles. Manipulating these angles allows the emulation of the camera's trajectory around the sphere, all while ensuring its gaze remains centered on the 3D mesh.

The compiled dataset encompasses a range of 3D meshes, RGB images captured from varied perspectives, and their associated depth maps. It is essential to note that the complexity of a structure often dictates the volume of data needed. The richer and more intricate the structure, the more comprehensive the dataset should be. This dataset is pivotal for training a neural network specifically designed for depth estimation of a particular structure, where the network takes an RGB image as its input and produces a depth map as its output. Subsequent sections will provide more details of the depth estimation network.

2.3. Displacement measurement

To measure displacement, a vibration video is first captured using a camera. This video is then processed through a segmentation model to remove the background and fed into a key point detector and tracker. Following the segmentation, the individual frames are further processed by the monocular depth estimation model to produce depth maps for each of them. Using these depth maps, along with the camera's intrinsic properties, a 3D surface point cloud for each frame is reconstructed. Then, the key points that have been tracked in each frame are used to accurately identify and mark their positions within the corresponding 3D point clouds. By pinpointing the 3D positions of the key points in each point cloud, the associated displacements can be accurately determined. The pipeline of the measurement system is shown in Fig. 5.

2.3.1. Depth estimation

A neural network [27] is employed to infer the depth map for every frame within the video sequence. While this network has previously been applied for 3D displacement measurement as detailed in [17], its accuracy in displacement determination remains restricted. Designed as an encoder-decoder structure, this network aims to predict a depth map from a solitary RGB input image. The encoder methodically reduces the image to a condensed latent representation. In contrast, the decoder works to upsample this representation back to the original input dimensions.

Deep convolutional neural networks (CNNs) effectively extract image features, discerning the relationship between input and output. As networks go deeper, performance can decrease due to gradient issues. To overcome this, ResNet [28] is employed in the model, which emphasizes residual learning and ensures stability in very deep networks. The ResNet50, pre-trained on ImageNet [29], serves as the encoder for our model, ensuring robust feature extraction. In the proposed architecture, the limitation of employing high-level semantic features, which inherently leads to coarse predictions, is addressed through a sophisticated decoder design. This decoder strategically combines high-level features, rich in semantic information, with low-level, edge-sensitive features to refine the feature maps back to their original resolution. This approach employs a progressive refinement strategy initiated by upsampling the

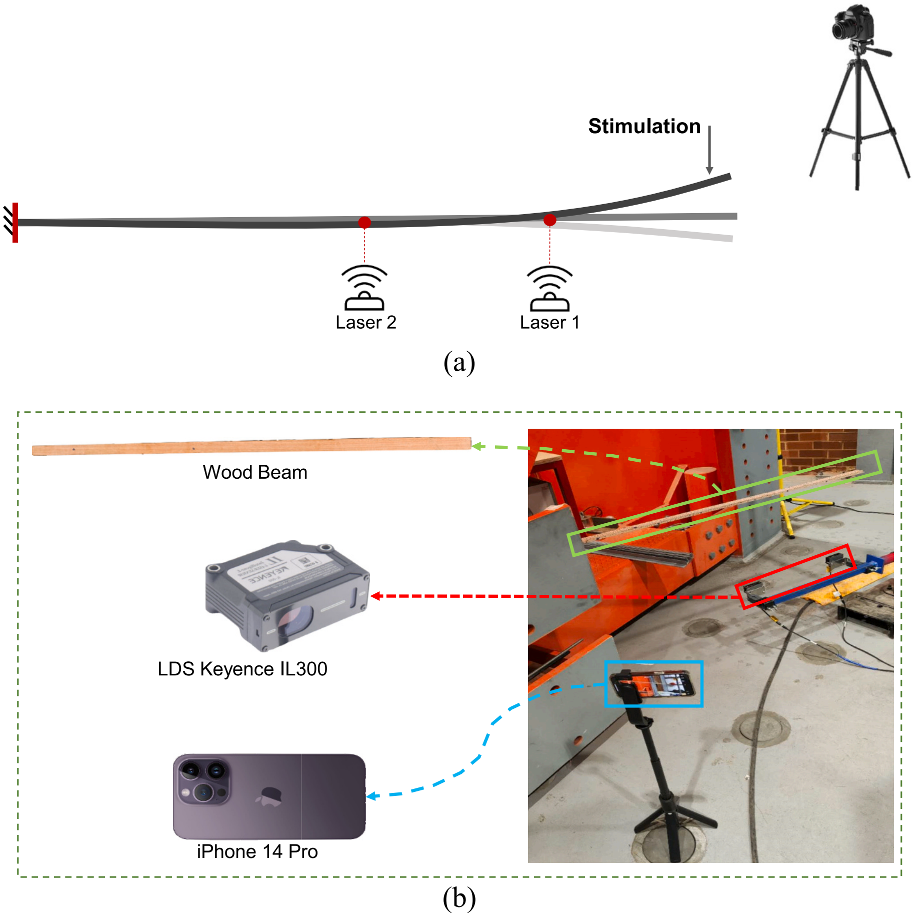


Fig. 7. Experimental test set-up: (a) schematic diagram of the experimental setup; and (b) on-site diagram of the experimental setup.

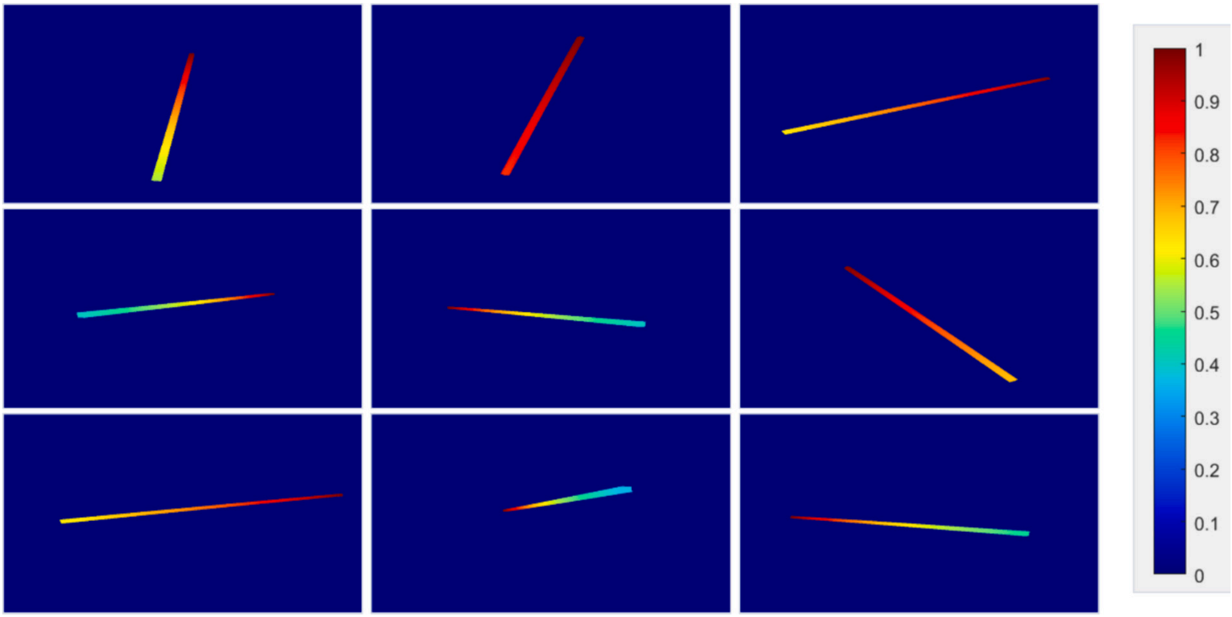


Fig. 8. Example depth maps sampled from training dataset.

last feature group produced by the encoder, enhancing detail recovery in the depth maps. Critical to this process is the utilization of residual convolution blocks, as highlighted in existing literature [30], which facilitate efficient gradient propagation from high-level to low-level layers. This is achieved via both short-range and long-range residual connections, ensuring a comprehensive integration of features across the network. Within this architecture, feature maps from designated encoder layers are transferred through a residual convolution block

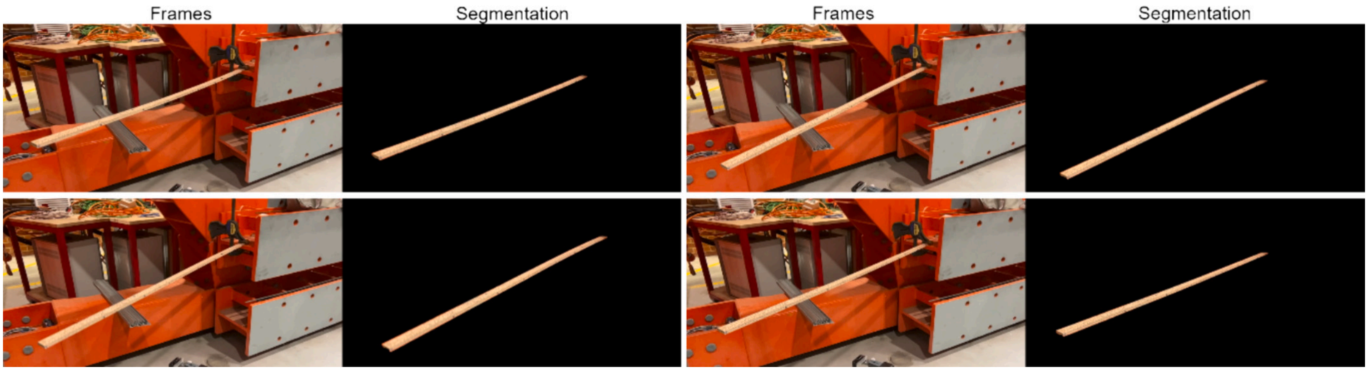


Fig. 9. Example segmentation masks of the wooden beam generated by TAM.

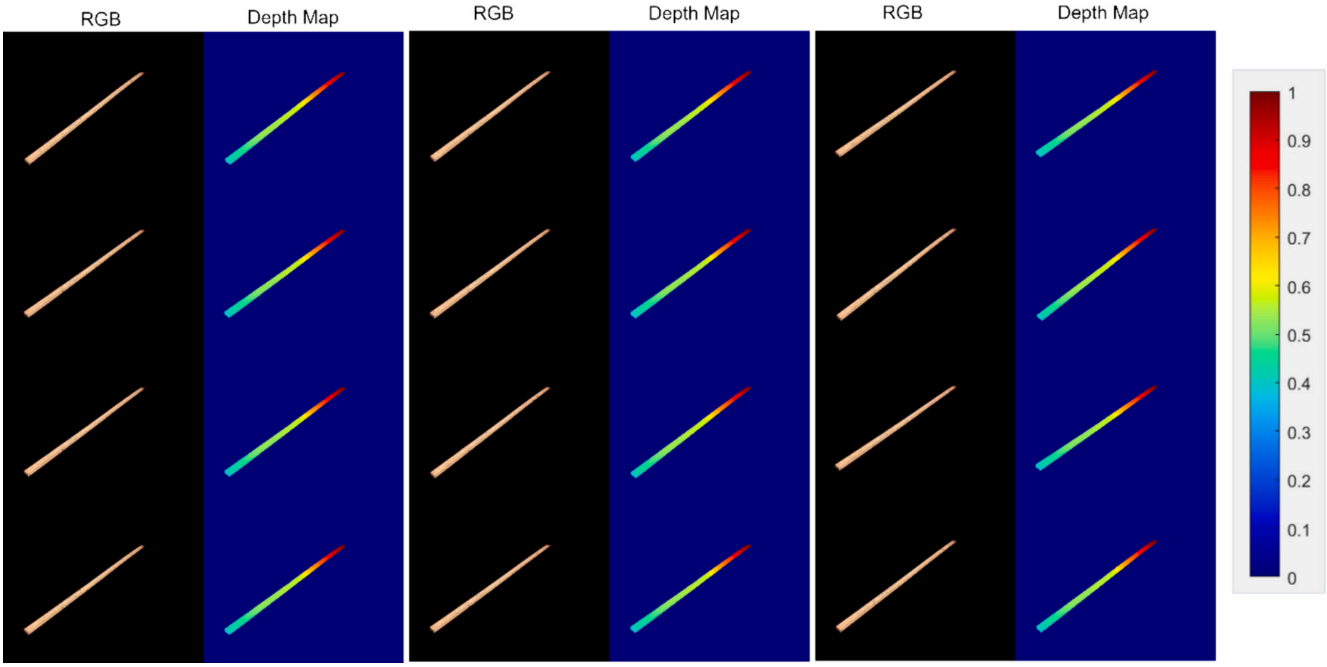


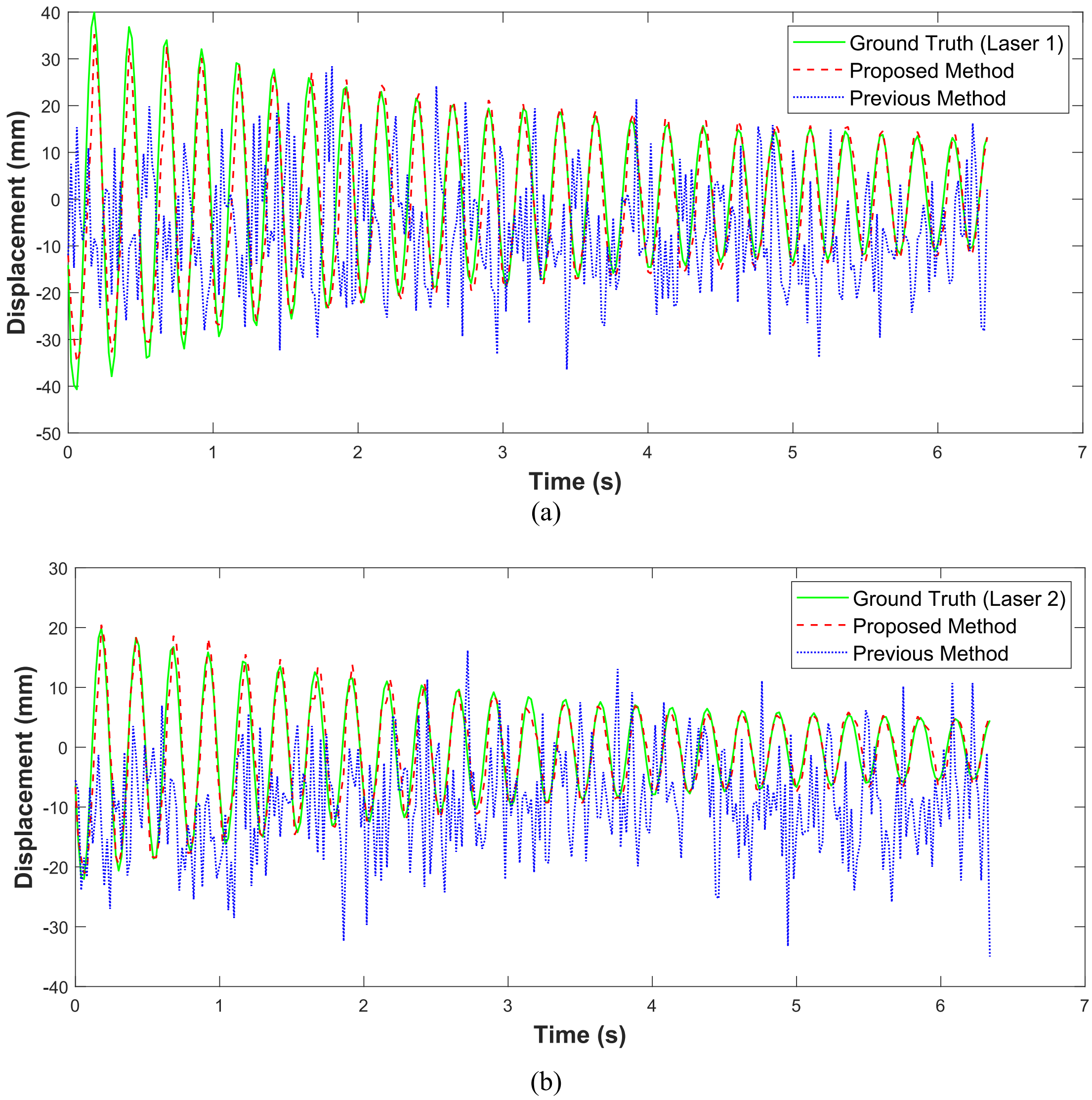
Fig. 10. Wooden beam depth map's samples as predicted by the fine-tunned depth estimation neural network.

within each fusion module. Prior to fusion, these maps undergo a transitional convolution layer, which standardizes the channel dimensions, facilitating seamless integration. The integration process involves merging these adjusted feature maps with those generated by the preceding feature fusion module, accomplished through a summation operation. Subsequently, an upsampling operation elevates the resolution of these fused maps to match that of the succeeding input layer. The final stage of the decoder incorporates an adaptive output module, comprising dual convolution layers and a bilinear interpolation layer, dedicated to outputting the final depth map.

Distinct from the works in Ref. [17], this paper introduces two primary modifications. Firstly, the availability of a high-quality dataset eliminates the need for the network to estimate affine-invariant depth. Instead of leveraging affine-invariant loss functions, the system now employs Chamfer loss and Mean Square Error loss to estimate scale-invariant depth. Data augmentation ensures scale-invariance of the depth maps in our dataset. Given that the cameras rendering the images and depth maps are positioned on a sphere with consistent intrinsic parameters, it is essential for the model to remain resilient to varying distances. In real-world applications, camera types and placements often differ. To prevent the model from overfitting to specific distances, random zoom in/out is used as a data augmentation strategy. This process modifies the perceived size of the object within the image and depth map, but the depth value range remains consistent. This strategy ensures that the model focuses on the object's inherent characteristics rather than its apparent size in the frame, steering the model to prioritize the object's shape over its size. Fig. 6 shows the zoom in/out augmentation. Secondly, observations during training revealed the auxiliary branch [27,31] for training on the generated dataset to be redundant, leading to its removal from the model.

Depth maps have limitations in out-of-plane displacement measurements. They solely offer the distance between 3D points and the camera. If the camera's principal axis is not aligned perpendicularly to the object, accurate displacement measurements become difficult. One possible approach is to estimate the relative positioning of the structure (world coordinate system) and the camera (camera coordinate system), but this often introduces significant inaccuracies. Real-world constraints, such as assessing tall structures, make it even more challenging to precisely determine this relative positioning between the camera and the structure. Hence, instead of the camera's filming position being meticulously adjusted or the relative positioning between the camera coordinate system and world coordinate system being estimated, the 3D point cloud of the structure is directly reconstructed using depth maps.

With the 3D point clouds in hand, even with only the surface points,



(a)

(b)

Fig. 11. Comparison of out-of-plane displacement time histories of wood beam: (a) Vision based method vs. Laser 1; (b) Vision based method vs. Laser 2.

the transformation matrix of the point cloud from the camera coordinate system to the desired world coordinate system can be estimated. This is attributed to the nature of civil structures, as their foundation is generally grounded, allowing points near the ground to serve as fixed points or the origin of the coordinate system. Once each frame's point cloud has been generated, reference points (at least 3 non-collinear ones) at fixed area can be chosen from a frame to establish the world coordinate system. If three non-collinear points are elusive, gravitationally defined normal can be adopted as alternatives. These points can then be matched with corresponding points across frames using the point tracking method, which will be introduced in the next section. Through this method, transformation matrices, encompassing rotation and translation between the camera and world coordinate systems, can be estimated.

Using Eq. (4), the 3D surface point cloud can be reconstructed once the depth map is estimated, combined with intrinsic parameters. The 3D surface point cloud represents the 3D location of every pixel. In essence, Eq. (4) back-projects each 2D pixel into its corresponding 3D point within the camera's coordinate system. A distinct feature of this measurement system, in contrast to many contemporary displacement measurement methods, is the absence of camera calibration requirements. The reason for this simplification is the scale-invariance of the estimated depth maps, allowing for the direct use of camera parameters from the rendering process of images and depth maps.

2.3.2. Key point detection and tracking

In order to find the dynamic movement of the measured structure, key points are used to represent the structure. Key points refer to specific pixels that have significant appearance, such as corner points [32]. These key points often help in measuring in-plane displacement. To figure out this displacement, the first frame is analyzed to pick out these key points, while other points are ignored. The KAZE detector [33] is used for this detection. Once selected, the KLT tracker [34–36] monitors these key points across different frames to see how they move. By looking at the movement of these key points frame-by-frame, a path for each point can be mapped out. This path represents the in-plane displacement. This in-plane displacement trajectory tells how a key point moves within a video frame by frame. Meanwhile, by mapping the trajectory of the key points to their respective depth maps, variations in the depth of the key points across the video can be obtained.

As previously described, in this measurement technique each image pixel is mapped back to a 3D location. As a result, for any given video, a sequential set of 3D surface point clouds is generated, each corresponding to an individual frame. By identifying the key point's position in every frame, its location within each of these point clouds can also be determined. Through key point tracking, the 3D coordinates of each key

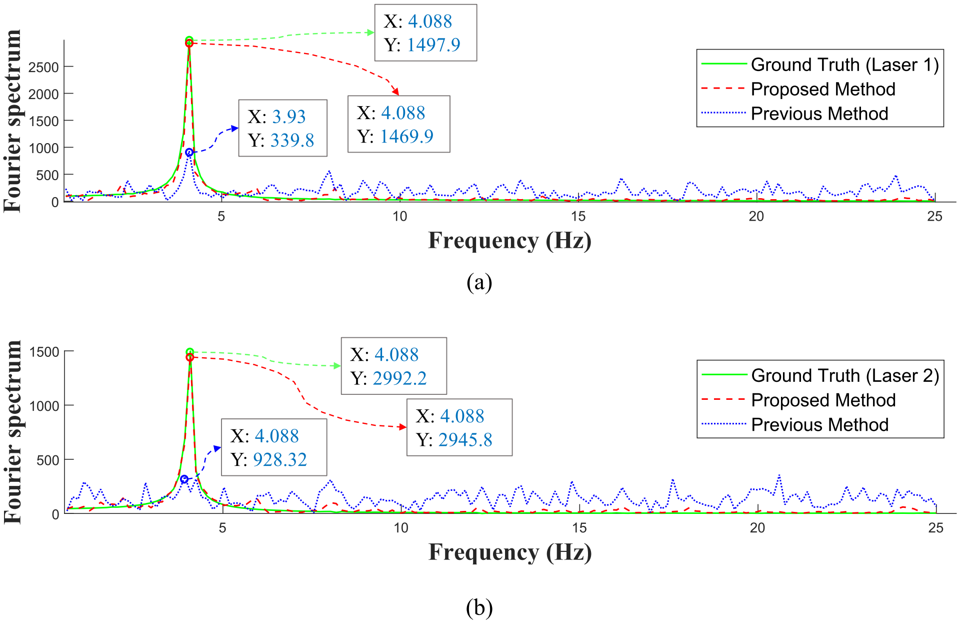


Fig. 12. FFT spectra of the measured displacement responses of wood beam: (a) Vision based method vs. Laser 1; (b) Vision based method vs. Laser 2.

Table 1
Cross-correlation coefficients and relative errors of key points at Lasers 1 and 2.

	Proposed Method		Previous Method [17]	
No.	**Corr. (ρ)**	**Relative Error. (%)**	**Corr. (ρ)**	**Relative Error. (%)**
G	1.0000	0.00	1.0000	0.00
Laser 1	0.9521	17.87	0.4249	106.16
Laser 2	0.9616	16.01	0.4720	94.56

point within the sequential point clouds can be determined. These coordinates can subsequently be transformed into the desired world coordinate system.

2.4. Object segmentation

It is infeasible and unnecessary to generate vast number of different backgrounds for civil structures in the custom-made dataset. Therefore, for testing video, a model called TAM (Tracking Anything Model) [37] is leveraged to extract the interest object and reject the background. With TAM, users can effortlessly segment the object of interest by a single click in the initial frame. The model then continuously tracks the chosen object across the video, from one frame to the next.

The TAM model is built upon a few advanced CV algorithms and architectures, each contributing to its robustness and efficiency. 1) Initialization with SAM (Segment Anything Model) [38]: The initialization phase employs the Segment Anything Model (SAM), which is designed for promptable image segmentation, comprising three integral components: an image encoder, a prompt encoder, and a mask decoder. The image encoder utilizes a pre-trained Vision Transformer (ViT) [39] adapted for high-resolution inputs, outputting a 16× downscaled, 64 × 64 image embedding. The prompt encoder is capable of handling sparse (points), dense (masks), and text prompts, each of which are embedded with positional embeddings, convolutional operations, and pre-trained text encoder CLIP [40]. To segment simple beam structures in this study, the user can easily select several points on the beam as the prompt. The mask decoder, designed for efficiency, employs a modified Transformer decoder block followed by a dynamic mask prediction head. It utilizes prompt self-attention and cross-attention mechanisms to update embeddings, ultimately mapping them to a binary segmentation mask through a dynamic linear classifier; 2) Tracking with XMem [41]: Following initialization, the model transitions to the tracking phase, which is managed by XMem. Given the initial object mask from SAM, XMem can track the object in the subsequent frames using ResNet-based architecture; 3) Mask Refinement with SAM: If the quality of the object

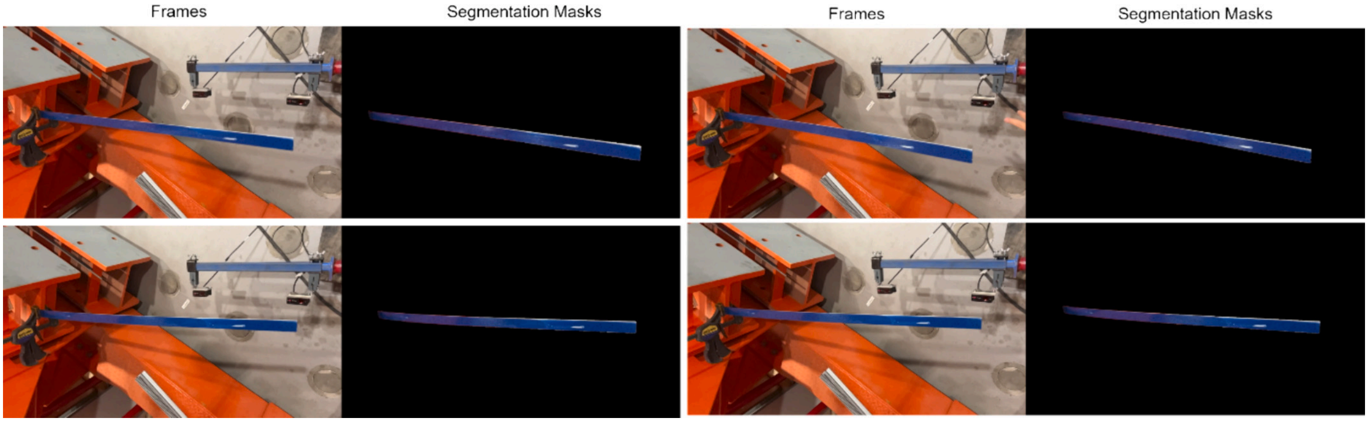


Fig. 13. Example segmentation masks of the aluminium beam generated by TAM.

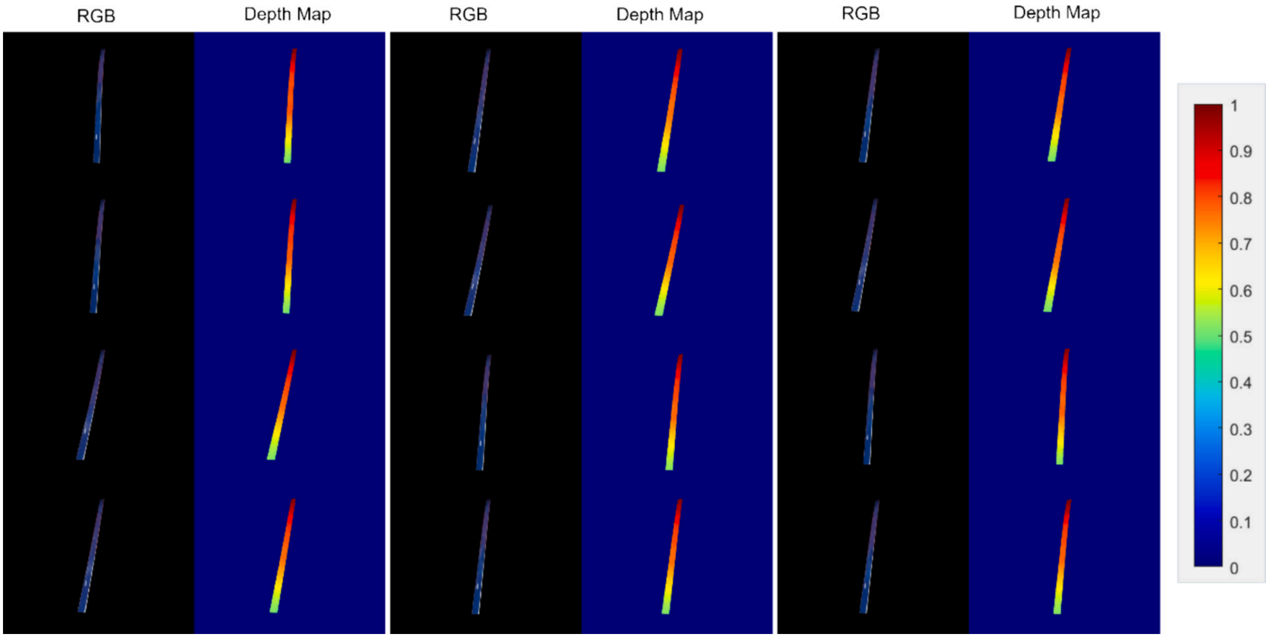


Fig. 14. Aluminium beam depth map's samples as predicted by the fine-tunned depth estimation neural network.

mask deteriorates during the tracking process, the model re-engages SAM for mask refinement. A feedback mechanism activates SAM, which uses dynamic hyperparameters like probe radius and affinity threshold for adjustments. The refined mask undergoes a quality check using IoU metrics to ensure effective refinement. This iterative refinement process allows the model to adapt to complex scenarios, such as occlusions or drastic changes in object appearance, thereby maintaining high tracking accuracy; 4) Real-time Correction: TAM incorporates a real-time correction mechanism that allows users to pause the tracking process and manually adjust the object mask. This feature is particularly useful for handling complex scenarios where automated tracking may fail due to occlusions or drastic changes in object appearance.

Using TAM, the object of interest can be precisely segmented in each frame. This not only streamlines the process by eliminating the need for background generation during dataset creation but also simplifies the training of the depth estimation network. Specifically, the absence of background allows the model to focus exclusively on the primary object, leading to potentially faster training convergence.

3. Experimental tests of beam structures

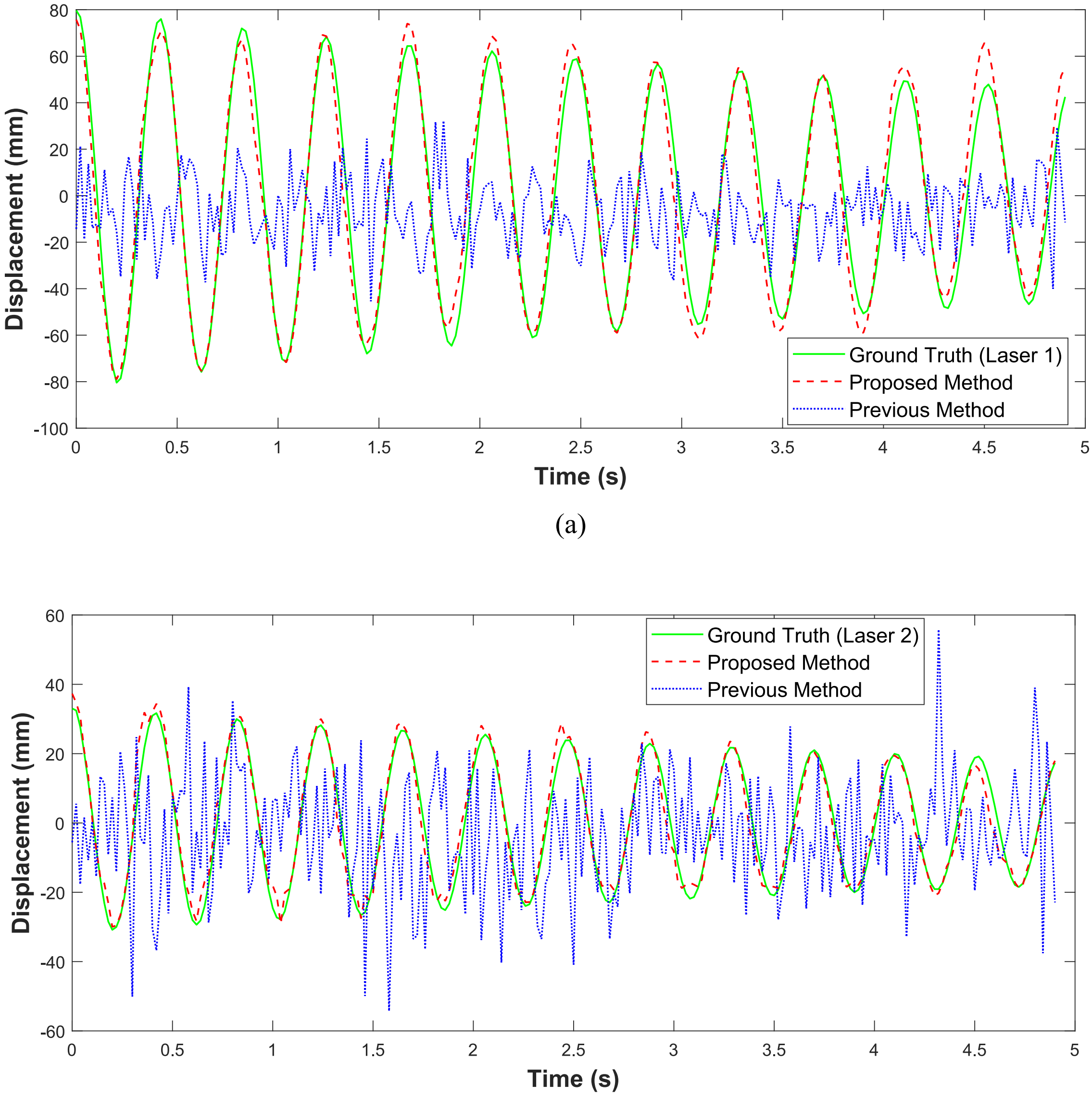
To validate the precision of the proposed approach for measuring out-of-plane vibration displacement in structural engineering, vibration experimental tests of two cantilever beams are conducted in a laboratory setting. The experimental tests utilized two distinct beams: one made of wood and the other of aluminium. The rationale for using two distinct beams lies in the differences in their different textures and vibration frequencies, despite both having the same cantilever mode shape. A reliable measurement system should be robust enough to accommodate these variations and provide consistent results across different materials and their associated modal behaviors.

3.1. Experimental set-up

Two distinct beams are employed as experimental specimens to evaluate the accuracy of the proposed measurement method across different materials and structural configurations. A wood cantilever beam, measuring 1200 mm in length, 30 mm in width, and 8 mm in thickness, and an aluminium beam, with dimensions of 1000 mm in length, 60 mm in width, and a thickness of 3 mm, were chosen. These beams were selected for their contrasting material properties and their common application in structural engineering. For the experiments, one end of each beam was securely anchored to a wall, creating a cantilever setup. An iPhone 14 Pro is employed to record the testing videos. Notably, rather than positioning the camera head-on to the structure's movement direction, it is deliberated angled. This deliberate orientation ensures the presence of out-of-plane movement, effectively testing the method's capability to measure such displacements. Fig. 7(a) provides a schematic representation of this setup. The recording parameters are set to a resolution of 1920 × 1080 and a capture rate of 155fps. For comparison and validation, two LDS Keyence IL300 sensors are placed on the structure's rear side to capture the 'ground truth' displacement data. Two lasers are employed to assess the full-field measurement capabilities of the system. If a good performance is observed at both laser-marked locations, full-field measurement is indicated. The detailed experimental configuration, including the placements of the sensors and other equipment, is shown in Fig. 7(b). During the experimental tests, the beam's free end is manually tapped to produce minor vibrations.

3.2. Dataset generation

To accurately measure the out-of-plane displacement using the depth estimation neural network, a comprehensive dataset is created by 3DGEN **[24]** , specifically tailored to the characteristics of a wood beam. A beam model is designed using particular dimensions that maintain a length, width, and height ratio. For the wood beam, this ratio is 1:0.025:0.00667, while for the aluminium beam, it is set at 1:0.06:0.003. This ratio is in accordance with the dimensions of the testing specimens. After the model is established, static loads are applied at ten distinct points along the beam span. These points start from one end of the beam (position 1) and shifts 0.05 L each time, in which L is the beam span length, continuing until the position reaches mid span of the beam. For each location, 100 different loading intensities are applied, with the intensities calibrated to ensure that the largest displacement of the beam remains within L/5 of the beam length. During the rendering process, for each deformed beam, 100 images and the corresponding depth maps are generated, captured from various unique views. In total, 200 K pairs of depth maps and images of each beam are generated for training the depth estimation neural network. Generating such dataset takes about 20 h. Some example depth maps are shown in Fig. 8.



(a)

(b)

Fig. 15. Comparison of out-of-plane displacement time histories of aluminium beam: (a) Vision based methods vs. Laser 1; (b) Vision based methods vs. Laser 2.

During training, the PyTorch framework was employed, adhering to a mini-batch strategy with a batch size set to 32. Optimization during training was facilitated through the Adam optimizer [42], a choice motivated by its adaptive learning rate capabilities, which has been configured with an initial learning rate of 0.001. To further refine the optimization process, the exponential decay rates for the moment estimates were meticulously set: 0.9 for the first moment and 0.999 for the second moment, aligning with common practices that balance responsiveness and stability in gradient updates.

3.3. Experimental results

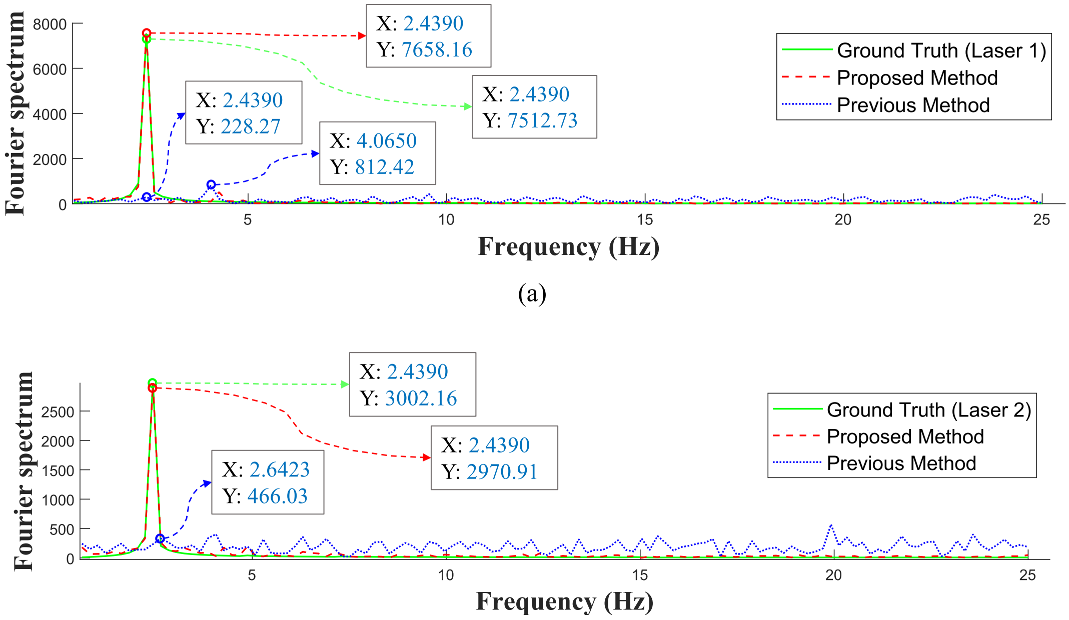
3.3.1. Wood beam experimental results

In Fig. 9, a series of segmented wooden beam images are presented next to their original RGB counterparts, showcasing the effectiveness of the Tracking Anything Model (TAM) in isolating and identifying structural elements.

Fig. 10 presents a selection of predicted wooden beam depth maps juxtaposed with their corresponding input RGB images. These visuals offer insights into the accuracy and capability of the depth estimation neural network, highlighting the intricate details captured.

Fig. 11(a) presents the out-of-plane displacement time history of an arbitrarily chosen key point in the area where Laser 1 is installed, comparing against the ground truth. Similarly, Fig. 11(b) illustrates the out-of-plane displacement derived from both the proposed system and the previous state-of-the-art (SOTA) monocular displacement measurement methods reported by Ref. [17]. Fig. 11 illustrates three separate displacement time histories: the green line represents the ground truth, the blue line corresponds to the results of the previous measurement method, and the red line shows the results of the proposed measurement system.

The results demonstrate that the proposed vision system delivers highly precise displacement measurements for wood beam, in line with those captured by physical sensors. When compared with the previous SOTA method, it is evident that the proposed measurement system significantly improves the accuracy. Fig. 12 illustrates the FFT spectrum of the measured displacements with different methods. It is evident that the proposed method aligns closely with the ground truth, demonstrating its effectiveness and precision in measuring out-of-plane displacements in terms of both vibration frequency and amplitude. In



(a)

(b)

Fig. 16. FFT spectra of the measured displacement responses of aluminium beam: (a) Vision-based method vs. Laser 1; (b) Vision-based method vs. Laser 2.

Table 2
Cross-correlation coefficients and relative errors of key points at Lasers 1 and 2.

	Proposed Method		Previous Method [17]	
No.	**Corr. (ρ)**	**Relative Error. (%)**	**Corr. (ρ)**	**Relative Error. (%)**
G	1.0000	0.00	1.0000	0.00
Laser 1	0.9810	19.43	−0.1249	194.32
Laser 2	0.9764	19.40	0.0376	124.39

contrast, the previous STOA method deviates significantly in the spectrum amplitude although it successfully captures the vibration frequency, highlighting its limitations and inaccuracies in out-of-plane displacement measurement.

Two evaluation metrics, namely the cross-correlation coefficient ρ and the relative error ϵ, are employed to assess the performance of the proposed approach. The parameters ρ and ϵ are defined as follows:

$$\rho = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{A_i - \mu_A}{\delta_A} \right) \left(\frac{B_i - \mu_B}{\delta_B} \right) \tag{5}$$

$$\epsilon = \frac{\|B_i - A_i\|}{\|B_i\|} \times 100\%, \tag{6}$$

where N represents the total number of observations in the displacement response's time history. The symbols B_i and A_i correspond to the ith displacement response from the ground truth and the proposed system, respectively. Additionally, μ_A and δ_A are the mean and standard deviation of A, while μ_B and δ_B are the mean and standard deviation of B.

The evaluation results for the displacements measured at locations with two LDS are detailed in Table 1. The displacement readings taken from the respective physical sensors serve as the benchmark or ground truth (denoted as G in the table). These readings have a relative error of 0 and a correlation coefficient of 1. The correlation coefficients and relative errors between the measured displacements at two laser sensor locations are also given in the table. The results clearly illustrate the capability of the proposed system to accurately measure the out-of-plane displacement. The cross-correlation coefficient between the measured displacement time histories from the laser sensor and the proposed methods at sensor location 1 is 0.9521, while at sensor location 2 is 0.9616. Furthermore, the accumulated relative errors are below 18%.

3.3.2. *Aluminium beam experimental results*

The experimental design for the aluminium beam maintains a high degree of similarity to the wooden beam's setup, with a key variation being the deliberate repositioning of the camera. This change is specifically intended to assess the algorithm's robustness concerning variations in camera placement, ensuring its adaptability and accuracy across different observational perspectives. Fig. 13 displays the RGB images captured by the camera alongside the segmentation masks of the aluminium beam.

Fig. 14 displays the RGB images captured by the camera alongside the depth maps inferred by the neural network.

Similar to Fig. 11, Fig. 15 shows three distinct displacement time histories measured along the aluminium beam under external stimulation at the free end. Fig. 15(a) displays the out-of-plane displacement time history of a randomly selected key point from the area where Laser 1 is positioned. In a similar vein, Fig. 15(b) depicts the out-of-plane displacement as determined by both the proposed system and the previous method at a location near Laser 2.

Fig. 16 showcases the FFT spectrum derived from aluminium beam displacements measured using the three methods. The displayed results clearly indicate that the proposed method closely mirrors the ground truth, highlighting its accuracy and reliability in gauging out-of-plane displacements.

Likewise, performance evaluation of the measurement system is carried out using the cross-correlation coefficient and the relative error. These metrics are presented in Table 2. At sensor location 1, the cross-correlation coefficient between the displacement time histories captured by the laser sensor and those from the proposed methods stands at 0.9810. For sensor location 2, this coefficient is 0.9764. Additionally, the total relative errors do not exceed 20%.

The paper refrains from detailing the in-plane displacement results. This is because the out-of-plane displacement, which is derived from the union of in-plane displacement and depth maps, acts as an indicator: its accuracy vouches for the precision of the in-plane displacement. The methodology reconstructs the structure's surface point cloud for each frame, facilitating displacement measurements in every conceivable direction. This 3D measurement prowess offers a comprehensive insight into the structure's movements across all dimensions. While the method can easily measure in-plane displacement, this essentially represents mere pixel movement in the videos. This pixel displacement cannot be transformed into any meaningful real-world displacement, due to the majority of the movement of the experiments are registered in the depth direction. Its relevance real-world analysis is minimal. Based on the results from two experimental tests, the proposed method consistently yields accurate displacement measurements. For both tests, the cross-correlation values exceed 0.95 and the relative errors remain below 20%. These findings highlight a significant improvement of the proposed measurement system over the previous method [17]. However, the relative errors observed from both wood beam and aluminium vibration tests are still relatively high. They can be attributed to the accumulation of errors throughout the measurement duration, as well as a combination of factors related to the equipment and computational constraints. Initially, there is a discrepancy in the data acquisition rates: the camera operates at 155.83fps, while the laser sensor at 200 Hz. Furthermore, due to computational limitations associated with the Training Anything [37] segmentation process, the camera's resolution must be reduced, resulting in videos being down-sampled to 50fps. Concurrently, the ground truth data from LDS was also asdjusted to 50 Hz. This down sampling process introduced synchronization challenges between the laser data and the video, potentially leading to the observed larger relative errors.

4. Conclusion

This paper proposes a novel system for out-of-plane displacement measurement leveraging monocular vision. The presented approach amalgamates a monocular vision-based depth estimation neural network with a novel 3D data generation technique, and an advanced large vision segmentation model to measure the full-field out-of-plane displacement of civil structures. When comparing with the previous state-of-the-art monocular depth estimation based measurement methods, the proposed system offers significantly superior measurement accuracy. Using the advanced synthetic data generation technique, data collection for training the neural network becomes significantly easier and time efficient. Two experimental tests are conducted to verify the effectiveness of the proposed measurement system. Results demonstrate that the proposed approach can measure the out-of-plane vibration displacement using monocular vision-based methods. Future research might explore the applicability of this vision system to more complex structures, like bridges and wind turbines. Additionally, integrating this measurement system into a UAV (Unmanned Aerial Vehicle) could also be a valuable avenue for exploration.

CRediT authorship contribution statement

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

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.

References

[1] M. Ghyabi, L.C. Timber, G. Jahangiri, D. Lattanzi, H.W. Shenton III, M.J. Chajes, M. H. Head, Vision-based measurements to quantify bridge deformations, J. Bridg. Eng. 28 (1) (2023), https://doi.org/10.1061/(ASCE)BE.1943-5592.0001960, pp. 05022010.

[2] Z. Ma, J. Choi, P. Liu, H. Sohn, Structural displacement estimation by fusing vision camera and accelerometer using hybrid computer vision algorithm and adaptive multi-rate Kalman filter, Autom. Constr. 140 (2022) 104338, https://doi.org/10.1016/j.autcon.2022.104338.

[3] L. Lu, F. Dai, A unified normalization method for homography estimation using combined point and line correspondences, Comput. Aided Civ. Inf. Eng. 37 (8) (2022) 1010–1026, https://doi.org/10.1111/mice.12788.

[4] J. Lv, M. Lv, J. Xiao, L. Wen, Q. Lou, A point tracking method of TDDM for vibration measurement and large-scale rotational motion tracking, Measurement 193 (2022) 110827, https://doi.org/10.1016/j.measurement.2022.110827.

[5] X. Pan, T.Y. Yang, Y. Xiao, H. Yao, H. Adeli, Vision-based real-time structural vibration measurement through deep-learning-based detection and tracking methods, Eng. Struct. 281 (2023) 115676, https://doi.org/10.1016/j.engstruct.2023.115676.

[6] M.A. Kuddus, J. Li, H. Hao, C. Li, K. Bi, Target-free vision-based technique for vibration measurements of structures subjected to out-of-plane movements, Eng. Struct. 190 (2019) 210–222, https://doi.org/10.1016/j.engstruct.2019.04.019.

[7] F. Chen, X. Chen, X. Xie, X. Feng, L. Yang, Full-field 3D measurement using multi-camera digital image correlation system, Opt. Lasers Eng. 51 (9) (2013) 1044–1052, https://doi.org/10.1016/j.optlaseng.2013.03.001.

[8] Y. Narazaki, F. Gomez, V. Hoskere, M.D. Smith, B.F. Spencer, Efficient development of vision-based dense three-dimensional displacement measurement algorithms using physics-based graphics models, Struct. Health Monit. 20 (4) (2021) 1841–1863, https://doi.org/10.1177/1475921720939522.

[9] X. Pan, T.Y. Yang, 3D vision-based out-of-plane displacement quantification for steel plate structures using structure-from-motion, deep learning, and point-cloud processing, Comput. Aided Civ. Inf. Eng. 38 (5) (2023) 547–561, https://doi.org/10.1111/mice.12906.

[10] S.W. Park, H.S. Park, J.H. Kim, H. Adeli, 3D displacement measurement model for health monitoring of structures using a motion capture system, Measurement 59 (2015) 352–362, https://doi.org/10.1016/j.measurement.2014.09.063.

[11] H. Yoon, H. Elanwar, H. Choi, M. Golparvar-Fard, B.F. Spencer Jr., Target-free approach for vision-based structural system identification using consumer-grade cameras, Struct. Control. Health Monit. 23 (12) (2016) 1405–1416, https://doi.org/10.1002/stc.1850.

[12] Y. Shao, L. Li, J. Li, S. An, H. Hao, Target-free 3D tiny structural vibration measurement based on deep learning and motion magnification, J. Sound Vib. 538 (2022) 117244, https://doi.org/10.1016/j.jsv.2022.117244.

[13] Y.F. Ji, C.C. Chang, Nontarget image-based technique for small cable vibration measurement, J. Bridg. Eng. 13 (1) (2008) 34–42, https://doi.org/10.1016/j.istruc.2023.105337.

[14] Y. Shao, L. Li, J. Li, S. An, H. Hao, Computer vision based target-free 3D vibration displacement measurement of structures, Eng. Struct. 246 (2021) 113040, https://doi.org/10.1016/j.engstruct.2021.113040.

[15] C.C. Chang, X.H. Xiao, Three-dimensional structural translation and rotation measurement using monocular videogrammetry, J. Eng. Mech. 136 (7) (2010) 840–848, https://doi.org/10.1061/(ASCE)EM.1943-7889.0000127.

[16] C. Sun, D. Gu, X. Lu, Three-dimensional structural displacement measurement using monocular vision and deep learning based pose estimation, Mech. Syst. Signal Process. 190 (2023) 110141, https://doi.org/10.1016/j.ymssp.2023.110141.

[17] Y. Shao, L. Li, J. Li, Q. Li, S. An, H. Hao, Monocular vision based 3D vibration displacement measurement for civil engineering structures, Eng. Struct. 293 (2023) 116661, https://doi.org/10.1016/j.engstruct.2023.116661.

[18] J. Sun, B. Peng, C.C. Wang, K. Chen, B. Zhong, J. Wu, Building displacement measurement and analysis based on UAV images, Autom. Constr. 140 (2022) 104367, https://doi.org/10.1016/j.autcon.2022.104367.

[19] X. Wang, C.E. Wittich, T.C. Hutchinson, Y. Bock, D. Goldberg, E. Lo, F. Kuester, Methodology and validation of UAV-based video analysis approach for tracking earthquake-induced building displacements, J. Comput. Civ. Eng. 34 (6) (2020), https://doi.org/10.1061/(ASCE)CP.1943-5487.0000928, pp. 04020045.

[20] R.I. Hartley, P. Sturm, Triangulation, Comput. Vis. Image Underst. 68 (2) (1997) 146–157, https://doi.org/10.1006/cviu.1997.0547.

[21] M. Visentini-Scarzanella, D. Stoyanov, G.Z. Yang, Metric depth recovery from monocular images using shape-from-shading and specularities, in: 2012 19th IEEE international conference on image processing, 2012, pp. 25–28, https://doi.org/10.1109/ICIP.2012.6466786.

[22] C. Tang, C. Hou, Z. Song, Depth recovery and refinement from a single image using defocus cues, J. Mod. Opt. 62 (6) (2015) 441–448, https://doi.org/10.1080/09500340.2014.967321.

[23] A. Raistrick, L. Lipson, Z. Ma, L. Mei, M. Wang, Y. Zuo, K. Kayan, H. Wen, B. Han, Y. Wang, A. Newell, H. Law, A. Goyal, K. Yang, J. Deng, Infinite photorealistic worlds using procedural generation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 12630–12641. https://arxiv.org/abs/2306.09310.

[24] Y. Shao, L. Li, J. Li, Q. Li, S. An, H. Hao, 3DGEN. https://github.com/YANDA-SHAO/3DGEN, 2023. (Accessed 22 March 2024).

[25] Trimesh. https://trimsh.org, 2019. (Accessed 22 March 2024).

[26] Blender - a 3D modelling and rendering package. http://www.blender.org. (Accessed 22 March 2024).

[27] W. Yin, J. Zhang, O. Wang, S. Niklaus, L. Mai, S. Chen, C. Shen, Learning to recover 3d scene shape from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 204–213. https://arxiv.org/abs/2012.09365.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.

[29] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: a large-scale hierarchical image database, in: 2009 IEEE conference on computer vision and pattern recognition, 2009, June, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848.

[30] G. Lin, A. Milan, C. Shen, I. Reid, Refinenet: Multi-path refinement networks for high-resolution semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1925–1934, https://doi.org/10.1109/CVPR.2017.549.

[31] Y. Liu, B. Zhuang, C. Shen, H. Chen, W. Yin, Auxiliary learning for deep multi-task learning, arXiv preprint arXiv:1909.02214, https://arxiv.org/abs/1909.02214, 2019.

[32] D. DeTone, T. Malisiewicz, 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, https://doi.org/10.1109/CVPRW.2018.00060.

[33] P.F. Alcantarilla, A. Bartoli, A.J. Davison, KAZE features, in: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7–13, 2012, Proceedings, Part VI 12, Springer, Berlin Heidelberg, 2012, pp. 214–227, https://doi.org/10.1007/978-3-642-33783-3_16.

[34] B.D. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, in: IJCAI'81: 7th International Joint Conference on Artificial Intelligence 2, 1981, August, pp. 674–679, https://doi.org/10.5555/1623264.1623280.

[35] J. Shi, Good features to track, in: 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1994, June, pp. 593–600, https://doi.org/10.1109/CVPR.1994.323794.

[36] C. Tomasi, T. Kanade, Shape and motion from image streams under orthography: a factorization method, Int. J. Comput. Vis. 9 (1992) 137–154, https://doi.org/10.1007/BF00129684.

[37] J. Yang, M. Gao, Z. Li, S. Gao, F. Wang, F. Zheng, Track anything: segment anything meets videos, arXiv preprint arXiv:2304.11968, https://arxiv.org/abs/2304.11968, 2023.

[38] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, G. Laura, T. Xiao, S. Whitehead, A. Berg, W. Lo, P. Dollar, R. Girshick, Segment anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026, doi: 10.48550/arXiv.2304.02643.

[39] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929, https://arxiv.org/abs/2010.11929, 2020.

[40] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning, 2021, July, pp. 8748–8763. PMLR, https://arxiv.org/abs/2103.00020.

[41] H.K. Cheng, A.G. Schwing, Xmem: Long-term video object segmentation with an Atkinson-shiffrin memory model, in: European Conference on Computer Vision, Springer Nature Switzerland, Cham, 2022, October, pp. 640–658. https://arxiv.org/abs/2207.07115.

[42] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, 2014 arXiv preprint arXiv:1412.6980. doi: 10.48550/arXiv.1412.6980.