Contents lists available at ScienceDirect

Engineering Structures

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

Monocular vision based 3D vibration displacement measurement for civil engineering structures

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

[a] School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, WA 6102, Australia
[b] Centre for Infrastructural Monitoring and Protection, School of Civil and Mechanical Engineering, Curtin University, Bentley, WA 6102, Australia

ARTICLE INFO

Keywords:
3D displacement measurement
Computer vision
Monocular vision
Deep learning

ABSTRACT

Vibration displacement of civil structures are crucial information for structural health monitoring (SHM). However, the challenges and costs associated with traditional physical sensors make displacement measurement difficult. In recent years computer vision (CV) techniques have been employed for measuring vibration displacement in civil structures. There has been a growing interest in CV-based three-dimensional (3D) displacement measurement, as it provides comprehensive information for structural health assessment. Most existing methods use multi-view geometry, requiring multiple cameras for depth measurement. This paper proposes a new system for measuring the 3D vibration displacement utilising a single camera. Instead of using multi-view geometry, deep neural networks are utilised to learn the depth of scenes from monocular images. Compared with the multi-view methods, the proposed 3D measurement system with monocular vision is more cost-effective and much more convenient to set up and use in practice, avoiding the complicated calibration and object matching between multiple cameras. Experimental tests are conducted in the laboratory to investigate the feasibility of the proposed system. Physical displacement sensors are equipped with the testing structure to provide the ground truth data. The results demonstrate that the proposed monocular 3D displacement system is able to produce reasonable 3D full-field displacement measurement, which makes monocular image based CV system a promising approach to achieve 3D displacement measurement, with its obvious advantages in cost and convenience compared to the traditional sensor-based or multi view CV-based methods.

1. Introduction

In the field of SHM, physical sensors, such as Linear Variable Differential Transformers (LVDT) and laser displacement sensors (LDS), are commonly used to measure the displacement of civil structures. Such sensors are very often difficult and costly to install and maintain. Furthermore, to install physical displacement sensors, stationary platforms are usually required to install them on. However, such fixed platforms can be difficult to set up in many in-field measurements. Therefore, accelerations are often measured instead, and displacement is obtained by double integration of the measured accelerations, which inevitably introduces some numerical errors especially when the baseline of the measured acceleration time history is difficult to be defined. On the other hand, computer vision (CV) based methods use cameras to measure the displacement. Using camera is much simpler than using displacement sensors, and the associated cost is often lower than using physical sensors.

Many studies have investigated vision based methods for displacement measurement [1-11]. Vision based displacement measurement can be divided into in-plane measurement and out-of-plane measurement. Camera imaging is a dimension reduction process, which projects 3D scenes into 2D images. Fig. 1 shows a monocular imaging system, where a point P in the 3D world coordinate system projects to a 2D pixel q in the image coordinate system. Assuming that the imaging plane is parallel to the 2D space spanned by Y_w and X_w, the spatial information of the 3D point P on Y_w and X_w directions can be easily derived from pixel q. However, the depth information along the depth direction (Z_c or Z_w) is lost. The in-plane displacement measurements refer to measuring the displacement in Y_w and X_w directions, while the out-of-plane displacement measurements aim to recover the displacement in the depth

* Corresponding authors at: School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Bentley, WA 6102, Australia (L. Li); Centre for Infrastructural Monitoring and Protection, School of Civil and Mechanical Engineering, Curtin University, Bentley, WA 6102, Australia (J. Li).

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.2023.116661
Received 16 January 2023; Received in revised form 18 July 2023; Accepted 23 July 2023
Available online 28 July 2023

direction. Measuring the in-plane displacement in both Y_w and X_w directions is termed as 2D displacement measurement, and measuring both the in-plane and out-of-plane displacement in X_w, Y_w and Z_w is termed as 3D displacement measurement.

In the last two decades, many CV based algorithms have been proposed for in-plane displacement measurement and some of them have been successfully applied in structural health monitoring [5,12]. 3D displacement measurements of structures provide a more comprehensive representation of structural performance. For example, bridges are critical structures that need regular monitoring for safety and stability. Measuring the 3D vibration displacement of bridges allows engineers to monitor bridge structural behaviour, detect any excessive movement or deformation, and assess the integrity of the structure. Heritage structures require careful monitoring to preserve their integrity. Measuring three-dimensional vibration displacement helps in monitoring the structural behaviour of these monuments and detecting any signs of damage.

For out-of-plane displacement measurement multi-view geometry based methods are usually used in the field of SHM [3,7,11]. Some methods affix specific type of markers on the structures as key points to enhance the performance of the multi-view 3D displacement measurement [7,11]. However, affixing artificial markers on structures is not always feasible, for example, putting artificial markers on large-scale infrastructure requires significant effort and is not feasible for inaccessible locations. Hence target-free methods [3,10] are more desirable to improve the practicability of vision-based displacement measurement. In target-free methods, artificial markers are replaced by key points automatically detected by algorithms such as Scale Invariant Feature Transform (SIFT) [13], Speeded Up Robust Features (SURF) [14], KAZE [15], SuperPoint [16] etc. Multi-view geometry methods reconstruct 3D scenes by triangulation [17], which requires the cameras' relative poses and matched key points to be properly estimated.

A general pipeline of the target-free multi-view geometry based 3D displacement measurement system is shown in Fig. 2. The pipeline can be separated into two parts: camera calibration and key point matching. The camera calibration part focuses on estimating the cameras' poses, and the key point matching part is for spatial and temporal key point matching. For camera calibration a planar pattern (e.g., a chessboard pattern) shown at different orientations is commonly employed which needs to be captured by multiple cameras [18]. Although camera calibration has been well studied in computer vision, its complexity is often underestimated. While it can often be done easily in a laboratory, this operation, can become very difficult, even unfeasible, for many in-field SHM applications. When the interested civil structures are located in some hard-to-reach places (e.g., rivers, reservoirs), requiring the existence of a chessboard pattern in multiple images is usually very difficult. For key point matching, key point detection algorithms are employed first to detect distinct pixels in an image as key points. The key point matching algorithm are then used to match the detected key points. For in-plane displacement measurement, the key points detected in the first frame of a video are chronologically tracked in the subsequent frames to locate the key points in every frame. The movements of civil structures are often slow, hence the change between consecutive frames is usually small. The chronological key point tracking can usually achieve a good accuracy. For the out-of-plane displacement measurement, the key points need to be detected and matched on images taken from different angles to reconstruct the depth using triangulation. The multi-view key point matching is more challenging than the key point tracking, if the key points in multi-view images have large movement. Several advanced key point matching methods, such as Superpoint [16] and Superglue [19] can be used to detect the matched key points. However, the key point matching may fail when the structure is texture-less. Furthermore, an extra control system is required to synchronize multi-view cameras. The high cost of multi-view camera systems is also a concern in practice.

Recently, monocular depth estimation (MDE) has attracted significant research interests and attention in computer vision. MDE is a technique that estimates the depth (or depth map) from a single image. The value of each pixel in the depth map represents the distance between its 3D object and the camera. With MDE, the out-of-plane displacement can be extracted from the depth map. Currently, there are generally three types of solutions to the depth estimation problem, that is, LiDAR (Light Detection and Ranging), shape-from-X [20,21] and deep learning [22-31]. LiDARs have been widely used for depth estimation in industry, e.g., for autonomous vehicles for depth estimation [32]. The expensive cost and high power consumption of LiDARs negatively impact the applications. Moreover, only the sparse depth map can be generated by LiDARs. Recovering the depth from a single image is an ill-posed problem, since the depth information is lost in the 3D to 2D projection of camera imaging. It is theoretically impossible to establish a mathematical model to back-map a 2D pixel to 3D. Although very challenging, pioneers estimated depth maps using depth cues. Such methods are named shapes-from-X. Tsai et al. [20] proposed a method to detect vanishing lines and the vanishing point, which assisted in the construction of a depth map. Tang et al. [21] introduced an approach to recover the depth map of a single image by using the principle of camera focus. The shape-from-X methods rely on hand-craft features and are usually not robust.

Currently, the state-of-the-art methods for MDE are deep learning approaches. They have surpassed traditional methods by a large margin. The deep learning based MDE can be divided into three categories according to different types of depth maps used, namely, metric depth, affine-invariant depth and relative depth. Scaling the ground truth depth (in engineering unit) with an arbitrary scaling factor can get metric depth. Eigen et al. [22] proposed a deep learning based network for MDE. It is a coarse-to-fine framework, where the coarse network learned the global depth on the entire image to obtain a rough depth map and the fine network learned the local features to refine the depth map. The training data are obtained from KITTI [33] and NYU [34]. These are

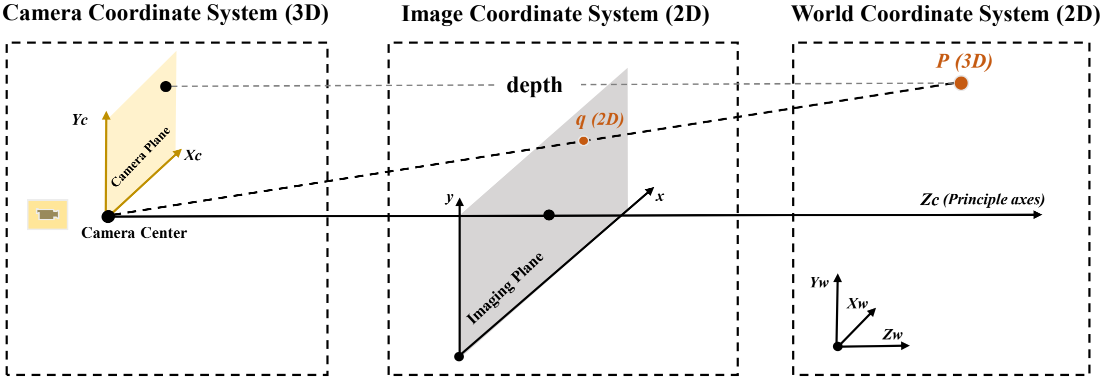


Fig. 1. The pinhole camera model of the monocular camera system.

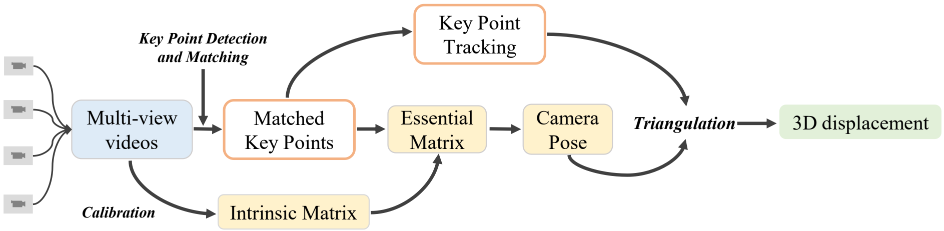


Fig. 2. The general pipeline of multi-view geometry vision-based 3D displacement measurement.

high quality metric datasets, since the depth maps are obtained by LiDARs and calibrated by stereo cameras. A scale-invariant loss in log space is designed and applied to compare the prediction with the ground truth. In 2018, Fu et al. [23] introduced a deep ordinal regression network (DORN) for MDE. Lee et al. [31] proposed a patch-wise attention mechanism to link the relationships among the neighbouring pixels in a local area. This network applies separate attention modules to each local patch. The separate attention maps are then merged as the output. The training dataset and loss function are the same as those used in [22], but much more promising results are achieved. Estimating metric depth requires high-quality datasets with ground truth, but existing metric depth datasets are limited in the diversity of scenes [25]. The commonly used metric depth datasets, for example NYU, consist mostly of indoor scenes with no human presence. The KITTI dataset involving many road scenes is mainly suitable for automatic driving tasks. The Make3D dataset [35] consists of mostly outdoor scenes of the Stanford campus. Although these datasets are of high quality and used widely for depth estimation, it is difficult to generalize the deep learning models trained from them.

To tackle the problem of model generalization, some studies collected stereo images from the web to generate diverse datasets. Large number of images are available online and they can be effectively used to improve the generalization ability of the learnt deep models [25]. However, metric depth map is hard to be obtained from online image pairs or videos since generally their camera intrinsic matrices are unknown. Disparity maps which refer to the pixel difference between a pair of stereo images are therefore used. The disparity is defined as $p_l - p_r = \frac{fb}{d_G}$, where p_r and p_l represent a pair of the corresponding points in the left and right images respectively, d_G is the ground truth depth, f is the focal length, and b is the baseline (i.e., the distance between the two cameras). In practice, the images taken by stereo cameras are sometime adjusted to release visual fatigue of viewers. Especially for 3D movies [36], film editors usually optimize the 3D footages by manipulating the optical centre of each camera, and the above equation becomes $p_l - p_r = \frac{fb}{d_G} + (O_l - O_r)$, where O_l and O_r are the optical centres of the left and right cameras respectively. The disparity maps of such stereo images are called affine-invariant depth maps. The generalization of the neural network models for affine-invariant depth estimation is better than those of the metric depth estimation models, due to the diversity of Internet images used for training. Many affine-invariant depth estimation methods are developed in recent years [25,26,28,36-38]. Xian et al. [25] presented a method that yields a disparity map dataset (ReDWeb) from about 40 k stereo images collected from the Flickr website. An encoder-decoder neural network was proposed based on ReDWeb, which has been adopted in many studies [29,38]. Wang et al. [36] proposed a Web Stereo Video Dataset (WSVD), which consists of over 7 k stereo videos collected from YouTube and Vimeo. The mean squared error (MSE) of the left and right images are calculated to ensure that the videos are stereo. Many videos have near-zero baselines, vertical disparities, inverted cameras, and other poor stereo characteristics. FlowNet2.0 [39] was utilised to remove these bad shots. A Normalized Multiscale Gradient loss [36] was employed to learn the affine-invariant depth. Affine-invariant depth allows MDE models to be trained on diverse datasets, which significantly increases the generalization of developed MDE models.

The field of computer vision has been the primary domain where MDE (Monocular Depth Estimation) models have been extensively developed and utilized. These models have also gained significant attention and found a popular application area in the realm of autonomous driving. They are yet to be employed for displacement measurement of civil structures. This paper proposes a novel deep learning based monocular vision measurement system to measure the 3D vibration displacement responses of civil structures. Different from the multi-view geometry based methods, only a single stationary camera is used to measure the 3D vibration displacement responses. The proposed measurement system is divided into two modules: the in-plane displacement measurement module and the out-of-plane displacement measurement module. To measure the in-plane displacement, numerous key points are detected using Superpoint [16] on the first frame and KLT tracker [40-42] is then applied to locate these key points on the subsequent frames. The in-plane displacement can be extracted from the movement of these key points. The out-of-plane displacement measurement module is based on the state-of-the-art monocular depth estimation (MDE) technique [29]. The affine-invariant depth estimation (ADE) neural network (NN) model is trained on diverse datasets, which enables the learnt deep NN model to have a good generalization capability. SHM requires converting the affine-invariant depth into engineering unit, and a metric depth recovering (MDR) neural network is used for this purpose. A flowchart of the proposed measurement system is shown in Fig. 3. To the best of the authors' knowledge, this is a pioneer study for the 3D vibration displacement measurement of civil structures using a monocular camera.

The rest of the paper is organized as follows: In Section 2, the methodologies for the proposed monocular vision based 3D vibration displacement measurement system are presented in details, focusing on the out-of-plane measurement. The performance of the proposed system is evaluated by two structural vibration tests. Experimental validations, performance comparisons and discussions are presented in Section 3. Conclusions are provided in Section 4.

2. Methodology

The proposed measurement system receives as input a structural vibration video taken by a single stationary camera and outputs the 3D displacement (in-plane-displacement and out-of-plane displacement) of many automatically detected key points. An example is shown in Fig. 4, assuming that a four-frame video is available, with four key points moving in the video. To extract the in-plane displacement, frame 0 is first fed into the key point detector which detects Point q_0 as a key point and discards other points. A deep learning based key point detector, Superpoint [16], is employed to detect key points on the subsequent frames and reject the other points. From frame 0 to frame 1, the key point q_0 moves to q_1. The KLT tracker [41,42] is used to track this movement, by generating a descriptor for the appearance features of point q_0 and find the point in Frame 1 (now Point q_1) based on the feature descriptor. The KLT tracker then generates a descriptor for key point q_1 to find it in frame 2 in a similar manner, and repeats the

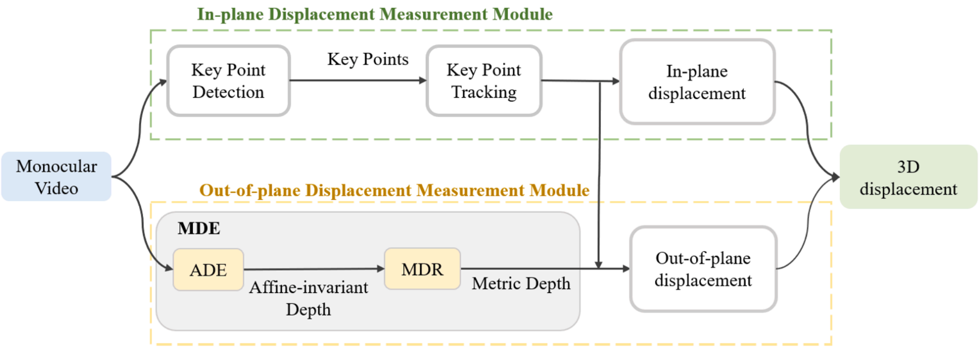


Fig. 3. The pipeline of the proposed monocular 3D displacement measurement system.

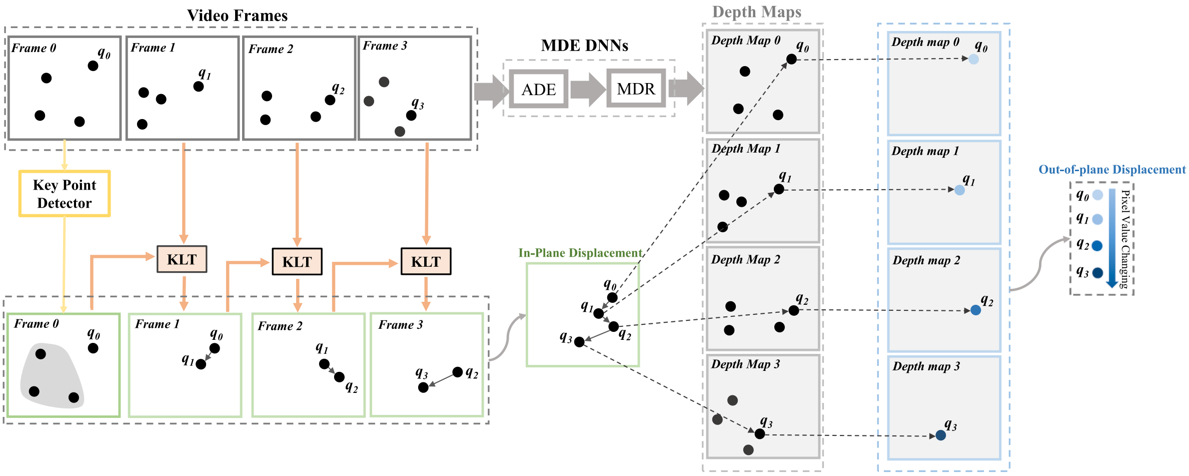


Fig. 4. Connection between the in-plane and out-of-plane displacement measurement modules.

describing and searching process until the last frame. In the end, a movement trajectory can be plotted for each key point based on its location in each frame, which forms the in-plane displacement measurement [43]. On the other hand, the out-of-plane displacement measurement module employs the MDE deep neural networks to estimate the depth maps for all video frames. The depth map has the same dimensions as the image frames, with the value representing the distance (depth) from the 3D object to the camera plane at each pixel

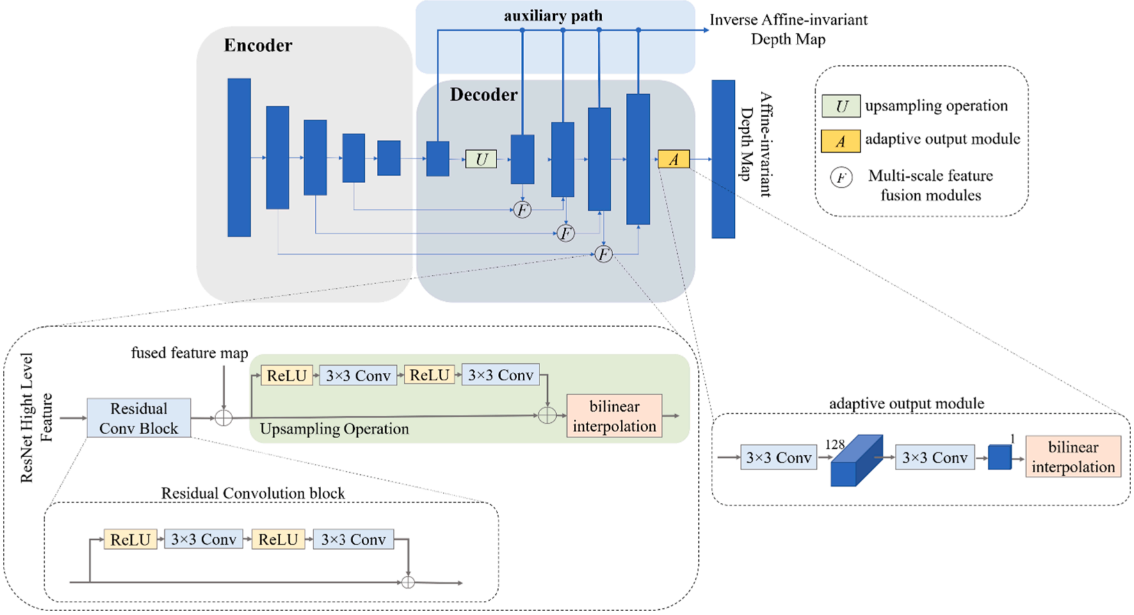


Fig. 5. ADE network consisted of three parts: an encoder for feature extraction, a decoder for depth map output, and an auxiliary path for inverse affine-invariant depth predicting.

location. To obtain the out-of-plane displacement, the key points extracted and matched in the in-plane displacement measurement module are also used as the key points in the out-of-plane displacement measurement. The changes in the pixel value of key point q in the depth maps at every frame form the out-of-plane displacement of Point q.

2.1. Affine-invariant depth estimation (ADE)

The definition of affine-invariant depth is expressed as

$$d_A = Sd_{iG} + \Delta_d, \tag{1}$$

where d_{iG} is the inverse ground truth depth, S is the scale factor and Δ_d is the shift of affine transformation. For SHM applications, the ground truth depth d_G with engineering unit (mm) needs to be recovered. ADE is a neural network [29] for estimating the affine-invariant depth map for each frame. An encoder-decoder structure is applied by the ADE to predict a depth map from a single RGB input image. The encoder progressively converts the image to a lower-dimensional latent representation, while the decoder upsamples it back to the input size. To improve the performance of the neural network training, an auxiliary path [44] is added to predict the inverse affine-invariant depth during training. The architecture of the ADE model is shown in Fig. 5.

Encoder. The image features can be effectively extracted by deep convolutional neural networks (CNNs), which ultimately learns the underlying mapping between the input and output. Ideally, the extracted feature becomes more and more powerful when the network goes deep. However, the network performance is dramatically decreased when the depth of the network increases to a certain level, due to gradient vanishing or exploding. ResNet [45] is used to tackle the problems that block the neural network from going deep. It explicitly allows the layers in the original CNN fitting a residual mapping rather than the underlying mapping between the input and output, which allows to train very deep networks. Owing to these advantages, the ResNet101 model pre-trained on ImageNet [46] is used as the encoder for feature extraction in the ADE model.

Decoder. Employing high-level semantic features leads to coarse predictions. A decoder that integrates high level features and low level edge-sensitive features is used to recover the feature maps to the original size. The decoder follows a progressive refinement strategy, which begins with an upsampling operation on the last group generated by the encoder. An existing study [47] showed that gradients from high-level layers can be efficiently propagated to low-level layers through short-range and long-range residual connections by residual convolution blocks. Feature maps from specific layers of the encoder are transferred by a residual convolution block on every fusion module, after a transitional 3 × 3 convolution layer is applied to adjust the channel number of feature maps. They are then merged with fused feature maps that are produced by the last feature fusion module via a summation operation. Finally, an upsampling operation is applied to generate feature maps of the same resolution as the next input. In the last layer, an adaptive output module that consists of two convolution layers and a bilinear interpolation layer is stacked to output the depth map. The feature maps of the decoder in each layer are directly shared with the auxiliary path to output the inverse affine-invariant depth, which jointly optimizes the model with the main ADE neural network. The auxiliary path is used in training, and discarded during inference.

Datasets. To train a model with good generalization capability, five datasets including Taskonomy [48], 3D Ken Burn [49], DIML [50], Holopix50K [37] and HR-WSI [51] are used. The datasets can be divided into three categories. The first one is the high-quality data, which are the Taskonomy dataset with the metric depths acquired by LiDARs and the 3D Ken Burn dataset with synthetic metric depths. The DIML dataset containing calibrated stereo images (metric depth) is considered the medium-quality data. The images in the Holopix50k and HR-WSI datasets are uncalibrated image pairs, hence are treated as low-quality data (with affine-invariant depth).

Loss function. Four loss functions are utilised to learn the affine-invariant depth when training the ADE network. They are image-level normalized regression (ILNR) loss [29], pair-wise normal regression (PNR) loss [29], ranking loss [26] and multi-scale scale-invariant gradient (MSSG) loss [52]. The ILNR loss is used to normalize the depth value in various datasets, which makes the depth stay within a certain range. The surface normal is an essential geometric feature for depth maps. The PNR loss uses surface normal to improve local and global geometric quality of depth maps. Ranking loss is used for the binary relation learning. Instead of training with fixed point pairs, an online sampling strategy resorting to explore the diversity of sampled point pairs [26] is applied in this study. Quality depth maps should have smooth gradient changes and big depth discontinuities. Image gradient which is usually used for edge detection can be used as an effective high-order geometric information to enforce the neural network to learn such feature. Thereby, the MSSG loss is used to achieve the above goals to enhance the quality of the predicted depth maps. According to the different qualities of the five datasets, different loss functions are applied. Depth maps in the high-quality datasets show clear and accurate planes and edges. For such datasets, the MSSG loss, ranking loss, ILNR loss and PNR loss are used. The MSSG loss, PNR loss and ILNR loss are used for the medium-quality datasets, in which the PNR loss is only used on planar regions since the regions around edges can be noisy. The low-quality datasets have unknown scale and shift factors, and the regions around edges and planes are noisy. Thus, only the ranking loss is used for the low-quality datasets. The ADE neural network is trained using stochastic gradient descent (SGD) with a batch size of 40. The initial learning rate is 0.02 for all layers, and a learning rate decay of 0.1 is applied. The images are evenly loaded from each dataset for each batch following a pervious study [28]. Table 1 lists the loss functions for each dataset used in this work.

2.2. Metric depth recovering (MDR)

The ADE network individually predicts the affine-invariant depth map for each frame. The factor S and shift Δ_d as shown in Eqn. (1) are different in different depth maps. Estimating two parameters for every depth map is very inefficient and difficult. Civil engineering structures usually consist of rigid body components, and the points on the rigid body components have fixed relations among them. When the whole structure vibrates, the relationship between the rigid points is generally fixed. Figure 6 shows an example of rigid structure. Point a and point b are a pair of points on a rigid structure. The absolute distance $r = d_{Gb} - d_{Ga}$ between point a and b is usually easy to be obtained by laser distance measuring devices. With the known r, when the affine-invariant depth can be converted into the metric depth $d_M = Sd_G$, the ground truth depth d_G can be recovered into engineering units.

A neural network MDR is used to predict Δ_d to convert the affine-invariant depth into the metric depth. MDR uses the Point-Voxel CNN (PVCNN) [53] network to learn Δ_d from 3D point clouds. For pinhole cameras, the back-projection formula from 2D pixel (u, v) to 3D point (x, y, z) is

$$\begin{cases} x = \dfrac{u - u_0}{f} d_M \\ y = \dfrac{v - v_0}{f} d_M \\ z = d_M \end{cases}, \tag{2}$$

Table 1
Matching between loss functions and datasets.

Dataset	Loss Function
Taskonomy [48]	MSSG, ranking loss, ILNR, PNR
3D Ken Burn [49]	MSSG, ranking loss, ILNR, PNR
DIML [50]	MSSG, PNR, ILNR
Holopix50K [37]	Ranking loss
HR-WSI [51]	Ranking loss

where (u_0, v_0) are the camera optical centre, f is the focal length, and d_M is the metric depth which is the inverse of d_{iM}. Adding a shift Δ_d to d_M will result in shape distortions of the point cloud. The pipeline of MDR training and inference is shown in Fig. 7. During training, the ground truth of the inverse metric depth is converted into affine-invariant depth by adding a ground truth depth shift Δ_d^*. The affine-invariant depth distorts the ground truth point cloud. MDR receives the distorted point cloud and predicts the depth shift Δ_d. The network is trained with the following objective function

$$\mathscr{L} = \min_{\theta} \left| \mathscr{M}\left(\mathscr{P}\left(u_0, v_0, u, v, f^*, \Delta_d^*, d_{iM}\right), \theta\right) - \Delta_d^* \right| \tag{3}$$

where $\mathscr{M}(\bullet)$ is the MDR network and $\mathscr{P}(\bullet)$ is the point cloud calculation, d_{iM} is the ground truth inverse metric depth and f^* is the ground truth focal length. During inference, the MDR network receives the affine-invariant depth predicted by ADE and estimates a depth shift Δ_d for converting the affine-invariant depth into inverse metric depth. To train the MDR network, 165,000 depth maps from ScanNet [54], Taskonomy and 3D Ken Burns are used.

2.3. In-plane displacement measurement module

The movements of key points are used to measure the in-plane displacements. Some pixels in images are ordinary or "boring". They are very similar to their neighbours, for instance, the pixels on a whiteboard. These pixels are hard to be used on downstream tasks such as point tracking, point matching, etc. Key points are some pixels that have a significant appearance, for example, the corner points. In the in-plane displacement measurement module, SuperPoint [16] is applied to detect key points. It is a self-supervised framework for learning key point detectors and descriptors, which are suitable for a large number of multiple-view geometry problems in computer vision. It achieves an excellent performance on many geometric computer vision tasks as well as structural displacement measurement tasks [3,43]. The SuperPoint neural network applies the Visual Geometry Group (VGG) convolutional neural network [55] as the backbone. A non-learned explicit decoder based on an Efficient Sub-Pixel Convolutional Neural Network (ESPCN) [56] is designed to upsample the feature map for resolution recovery. SuperPoint is trained on 80,000 wrapped MS-COCO datasets [57]. All training is conducted using PyTorch with mini-batch sizes of 32. Adam optimizer [58] is used during training with a default learning rate of 0.001 [16]. The exponential decay rates for the first and second moment estimates are 0.9 and 0.999, respectively. To locate the Superpoints on all the frames, the KLT tracker [40-42] is used. The KLT tracker detects the location of key points by comparing the neighbour pixels of the key points. An outlier-removing algorithm [59] is used to optimize the tracked key points. Note the key points detected in this module are also used as the key points for the out-of-plane displacement measurement.

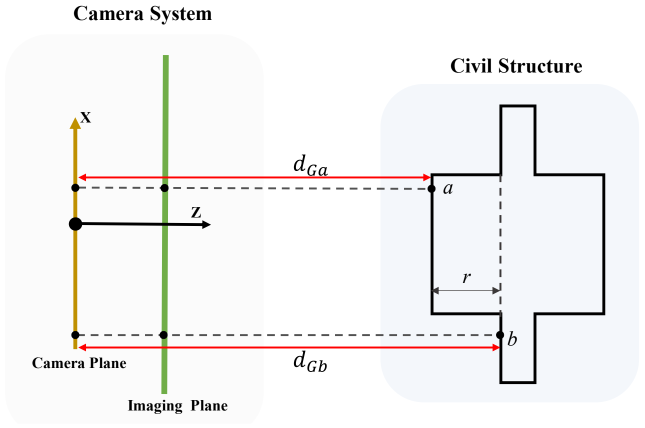


Fig. 6. An example of the engineering unit recovered by using the metric depth map.

3. Experimental validations

3.1. Experiment A

3.1.1. Experimental setup

This experiment aims to evaluate the performance of the out-of-plane displacement measurement module. A cantilever steel beam ($1500 \times 50 \times 10mm$) is installed on a shaker which provides the vibration displacement in the depth direction. A SONY PXW-FS 5 4 K XDCAM camera with a Sony E PZ 18 −105 mm F 4 GOSS Len is used for filming, which is facing the Z direction of the measured structure, as shown in the Top View in Fig. 8(a). The video resolution is 1920 × 1080 and the frame rate is 50*fps* (frames per second). The camera is placed 3.32*m* away from the structure, which is measured by a Bosch GLM 400 Laser Range Finder Distance Measurer. An LDS Keyence IL 300 is installed on the back of the structure to measure the displacements in the depth direction, which are used as the ground truth. A steel block that has known dimension in depth direction is mounted on the steel beam as a reference to convert the movement trajectory to the engineering unit. The experimental setup is shown in Fig. 8. Sinusoidal excitation with an excitation frequency of 2 Hz is applied to generate the vibrations in the z-direction.

3.1.2. Experiment A results

Ten Superpoint key points are detected for analysis within the area surrounding the location where the LDS sensor is installed at the back of the specimen. This allows for a direct comparison between our method and the ground truth measurements provided by the LDS. In Fig. 9, the out-of-plane displacement trajectory of an arbitrarily chosen key point is presented against the ground truth, where the ground truth is shown in orange, and the one measured by the proposed system is presented in blue. Figure 9(a) is the time history from 0 s to 20 s, while Fig. 9(b) is the zoomed-in view of the period from 10 s to 12 s.

The vibration displacement responses measured in the time domain are then converted to the frequency domain by Fast Fourier Transformation (FFT) to evaluate the performance of the proposed system in another view, as shown in Fig. 10. The dominated frequency component measured by both the proposed system and LDS is 1.998 Hz. A second frequency of 3.996 Hz is observed as shown in Fig. 10(b) by the proposed measurement system while a frequency of 3.546 Hz is observed as shown in Fig. 10(a) by LDS. It should be noted that the vibration of the structure is dominated by the applied sinusoidal excitation at 2 Hz. Therefore, both the proposed vision system and the LDS have captured the most significant vibration energies at 2 Hz as shown in Fig. 10(a) and (b). The frequency component of 3.996 Hz as shown in Fig. 10(b) is the second order harmonic frequency of the applied excitation. The observed frequency at 3.546 (the analytical natural frequency is 3.63 Hz based on finite element analysis with the dimensions and material properities of the cantilver beam) in Fig. 10(a) is the fundamental natural frequency of the cantilever beam, though its energy is very small. This is also the reason why it could not be identified in Fig. 10(b), since the vibration energy is small. To further investigate the frequency components of the obtained dynamic responses, the Fourier spectrum is converted to a base-10 logarithmic scale, with the one from LDS shown in Fig. 10(c) and the one from the proposed system shown in Fig. 10(d). It can be observed from Fig. 10(c) that there is a small peak at 3.996 Hz, which coorsponds to the second order harmonic frequency component in Fig. 10(b). However, it should be noted that in this testing, the obtained dynamic displacement responses are dominated by the applied sinusoidal excitation at 2HZ.

Two numerical evaluation parameters, cross-correlation coefficient ρ and relative error $\in$, are used to evaluate the performance of the proposed system on all selected 10 key points to ensure that the out-of-plane displacement measurement module has a good performance for all key points. ρ and $\in$ are defined as:

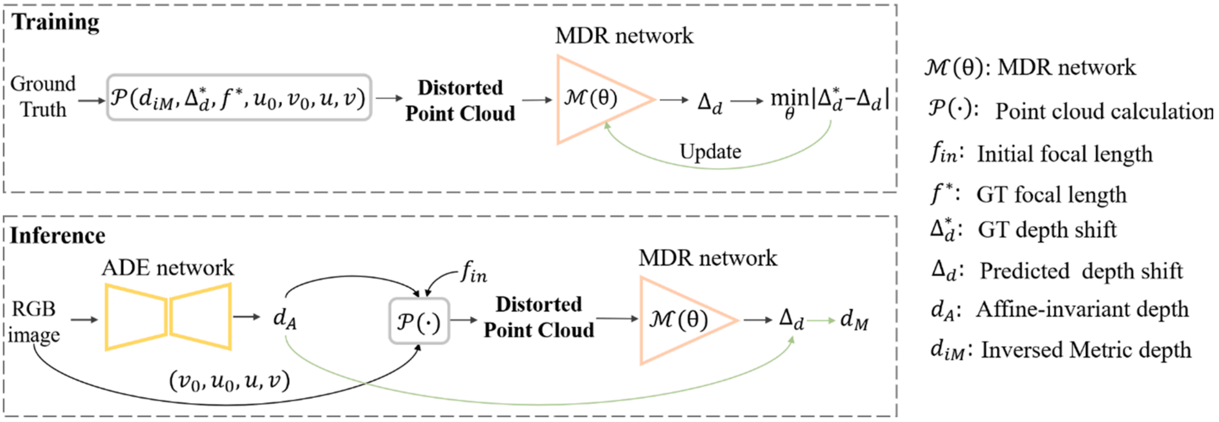

Training
Ground Truth
$\mathcal{P}(d_{iM}, \Delta_d^*, f^*, u_0, v_0, u, v)$
Distorted Point Cloud
MDR network
$\mathcal{M}(\theta)$
Δ_d
$\min_\theta|\Delta_d^* - \Delta_d|$
Update
Inference
ADE network
RGB image
d_A
f_{in}
$\mathcal{P}(\cdot)$
Distorted Point Cloud
MDR network
$\mathcal{M}(\theta)$
Δ_d
d_M
(v_0, u_0, u, v)
$\mathcal{M}(\theta)$: MDR network
$\mathcal{P}(\cdot)$: Point cloud calculation
f_{in}: Initial focal length
f^*: GT focal length
Δ_d^*: GT depth shift
Δ_d: Predicted depth shift
d_A: Affine-invariant depth
d_{iM}: Inversed Metric depth

Fig. 7. The pipeline of MDR method.

$$\rho = \frac{1}{N-1}\sum_{i=1}^{N}\left(\frac{A_i-\mu_A}{\delta_A}\right)\left(\frac{B_i-\mu_B}{\delta_B}\right) \tag{4}$$

$$\epsilon = \frac{\|B_i - A_i\|}{\|B_i\|} \times 100\% \tag{5}$$

where N is the number of observations in the time history of the displacement responses, B_i and A_i denote the ith displacement response of the ground truth and that obtained from the proposed system, respectively; μ_A and δ_A, μ_B and δ_B are the mean values and standard deviations of A and B, respectively. The numerical evaluation results are presented in Table 2. The obtained vibration displacement responses of the fifth key point (arbitrarily chosen) are shown in Figs. 9 and 10. It can be observed that the proposed out-of-plane displacement measurement module works stably on all ten key points, indicating that it could be used for full-field displacement measurement. The average cross-correlation and relative error of the out-of-plane vibration displacement measurement are 0.82 and 55.405%, respectively. This relatively large error can be attributed to the accumulated errors over the entire measurement duration of 20 s. Although the relative error is quite high, the cross-correlation value reaches 0.82. It should be noted that this is a pioneer work attempting to measure vibration displacement responses in 3D using a single camera, therefore metrics comparison is not available. Considering the significant challenges in measuring vibration displacement in the depth direction using a single camera, the accuracy achieved by the proposed system is encouraging and promising.

3.2. *Experiment B*

3.2.1. *Experimental setup*

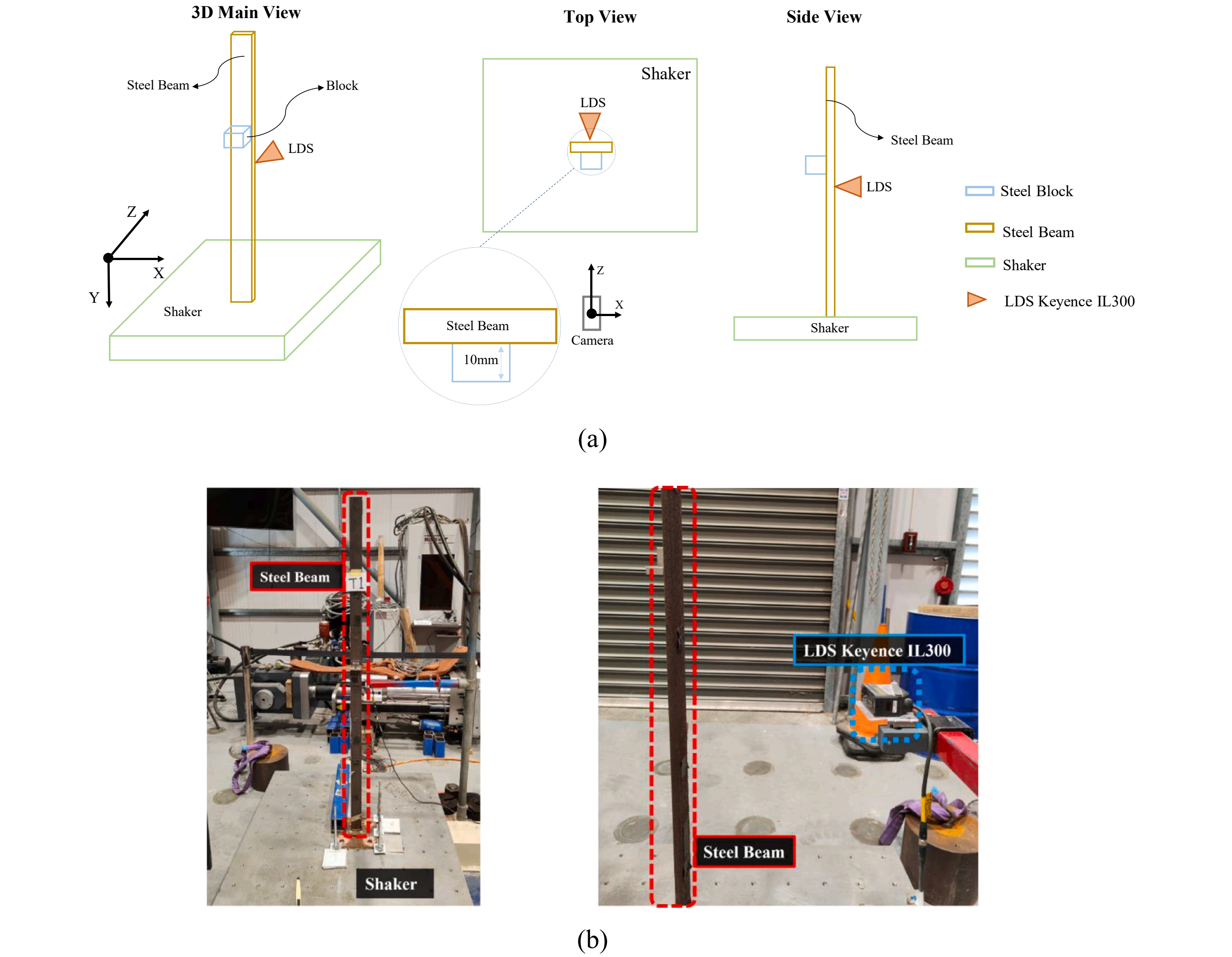
In this experiment, 3D vibration tests are conducted on a steel cantilever beam ($425 \times 50 \times 5mm$) to verify the performance of the proposed monocular system for 3D vibration displacement measurement. This test evaluates the performances of both the in-plane and out-of-plane displacement measurement. 3D vibrations are generated by two shakers: Shaker A and Shaker B. Shaker A is a bi-direction shaking table that provides excitations in the X (horizontal) and Z (depth) directions. Shaker B (APS 400 ELECTRO-SEIS long-stroke shaker) is fixed on shaker A for providing Y (vertical) direction excitation. The beam structure is fixed on shaker B. The combination of using two shakers generates 3D vibrations for the beam structure. The ground truth data of displacement responses are measured by four physical displacement sensors. The types and measurement directions of the sensors used are shown in Table 3. The camera and lens in this experiment are the same as those in experiment A. A remote controller is used to turn on and off the filming. The recording image resolution is 1920 × 1080 and the frame rate is 50*fps*. The camera is placed 1.76m away from the cantilever beam in the Z direction. Shaker A is employed to generate sinusoidal excitations with an amplitude of 3 mm and a frequency of 3 Hz in both the X and Z directions. Shaker B is controlled by adjusting the voltage of the shaker controller, which cannot provide the exact value of vibration displacement amplitude and frequency. The experimental setup of the shakers and cameras for Experiment B is shown in Fig. 11.

3.2.2. *Experiment B results*

The proposed in-plane and out-of-plane displacement measurement modules are used to measure the 3D displacements of 30 key points detected by SuperPoint. Fig. 12 shows the displacement response of an arbitrarily chosen key point measured by the proposed system against those measured by the physical sensors in X, Y and Z directions, where Fig. 12(a), (c) and (e) show the time history of displacement in the X-Y and Z directions from 0 s to 20 s, and (b), (d) and (f) show their respective zoom-in view. The results demonstrate that the in-plane module obtains very accurate displacement measurement against those measured by physical sensors. In depth Z direction, the accuracy of the displacement measurement from the out-of-plane module is obviously much lower, however they are still acceptable especially in terms of the frequency of the displacement. It can be observed that the measured value from the vision system is larger than ground truth in some peaks in Z direction. This can be attributed to the inherent nature of the neural network which processes each video frame independently. As the structure undergoes 3D movement, the captured image varies depending on the direction and magnitude of the structure's movement, such as moving towards or away from the camera, or moving right or left relative to the camera's perspective. Each frame presents a unique image with different spatial relationships and depth cues, which can affect the network's performance in estimating depth. Another observation is that the vision displacement measurement seems to perform better on the positive side than the negative side most of the time. A possible explanation is that the neural network estimates the depth map based on depth cues present in the scene. When the structure moves closer to the camera, it tends to occlude more objects in the scene, resulting in a decrease in available depth cues. It could become more challenging for the network to accurately estimate depth in such scenarios. In contrast, when the structure moves away from the camera, it may reveal more objects or more structural parts in the scene, providing more distinct depth cues for the network to leverage. This can lead to improved performance in depth estimation.

Again, the time domain displacement responses are transformed into the frequency domain by FFT. Figure 13 presents the frequency domain responses of the 3D displacement measured by the proposed system and the physical sensors.

It can be observed that the vibration frequencies measured in the X and Y direction match very well with the data from LVDTs which are used as the ground truth. For the Z-direction, the same response fre-



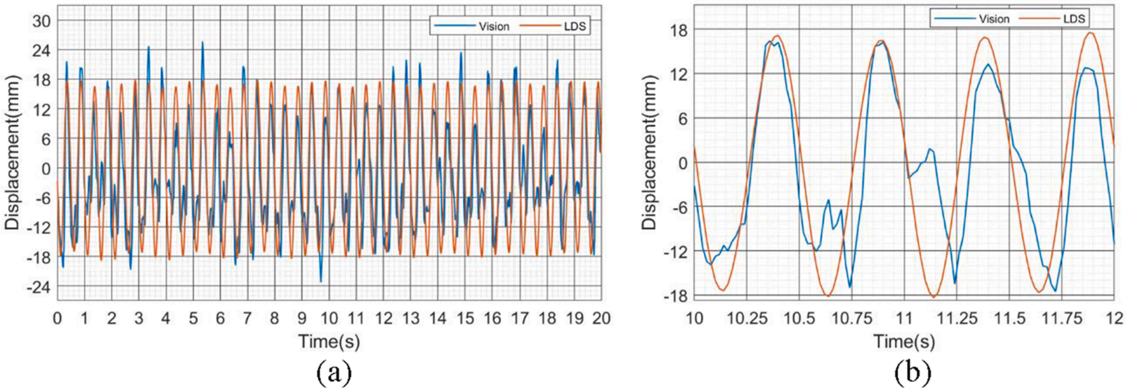
(a)

(b)

Fig. 8. Setup of Experiment A.

quencies of 2.997Hz and 5.994 Hz are also observed in the data from the proposed system and the ground truth. From the experimental results, it can be observed that the proposed vision system is able to achieve accurate measurement for in-plane displacement. The dominant frequencies of the out-of-plane displacement are also well measured, although the amplitude of the out-of-plane response and the corresponding Fourier spectrum amplitude have some differences.

Similar observation can be made in the log scale, as shown in Fig. 14 . The proposed system produces the vibration displacement in both the *X* and *Y* directions accurately, as shown in Fig. 14(a–d). In the *Z* direction,



(a)

(b)

Fig. 9. Out-of-plane displacement of one arbitrarily chosen key point using the proposed system against the ground truth measured by LDS. (a) Vision vs. LDS in Z direction; (b) A zoomed-in view of (a).

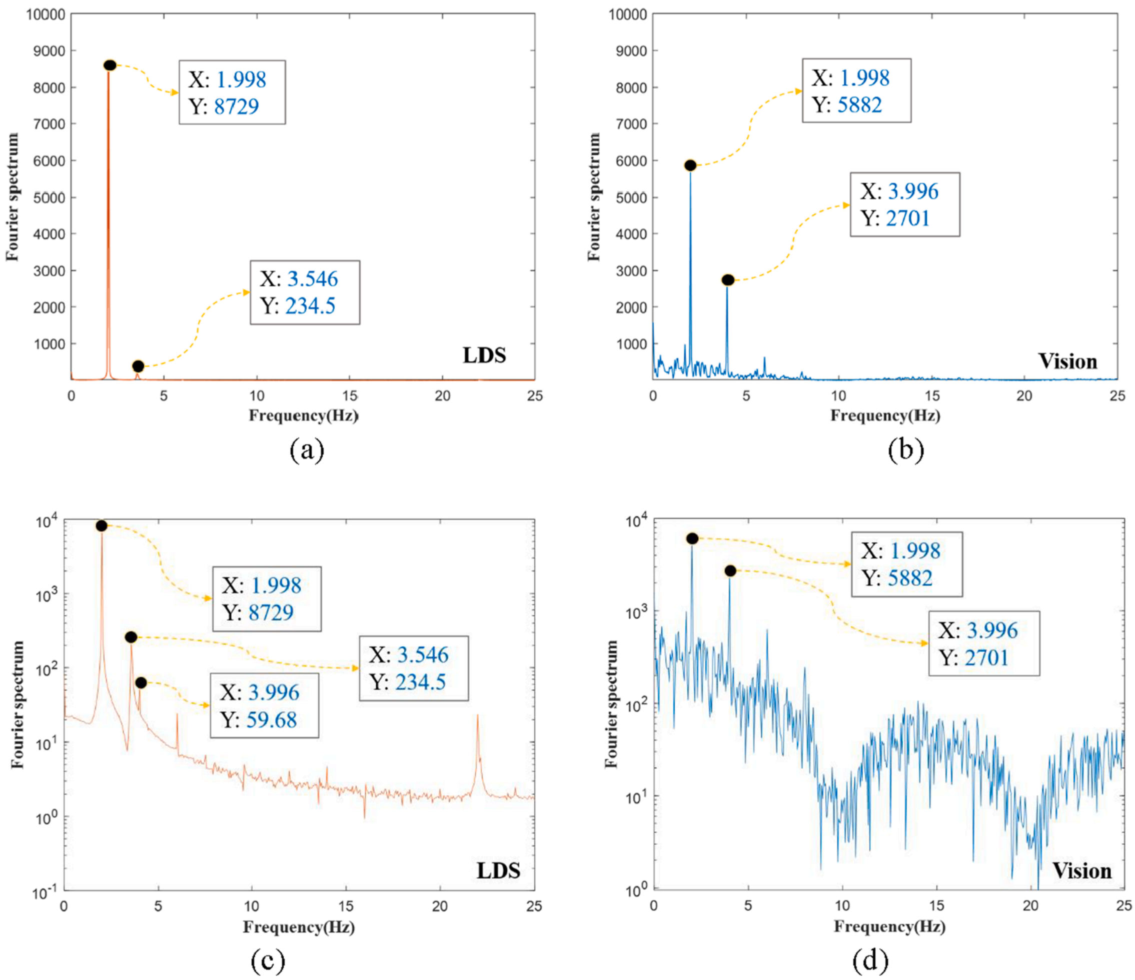


Fig. 10. FFT spectra of out-of-plane vibration displacements measured by the proposed system and LDS: (a) by LDS ; (b) by the proposed system; (c) in the logarithmic scale measured by LDS; (d) in the logarithmic scale measured by the proposed system.

the proposed system is able to detect the main features of the vibration responses, but with certain discrepancies in the amplitude, as shown in Fig. 14(e) and (f).

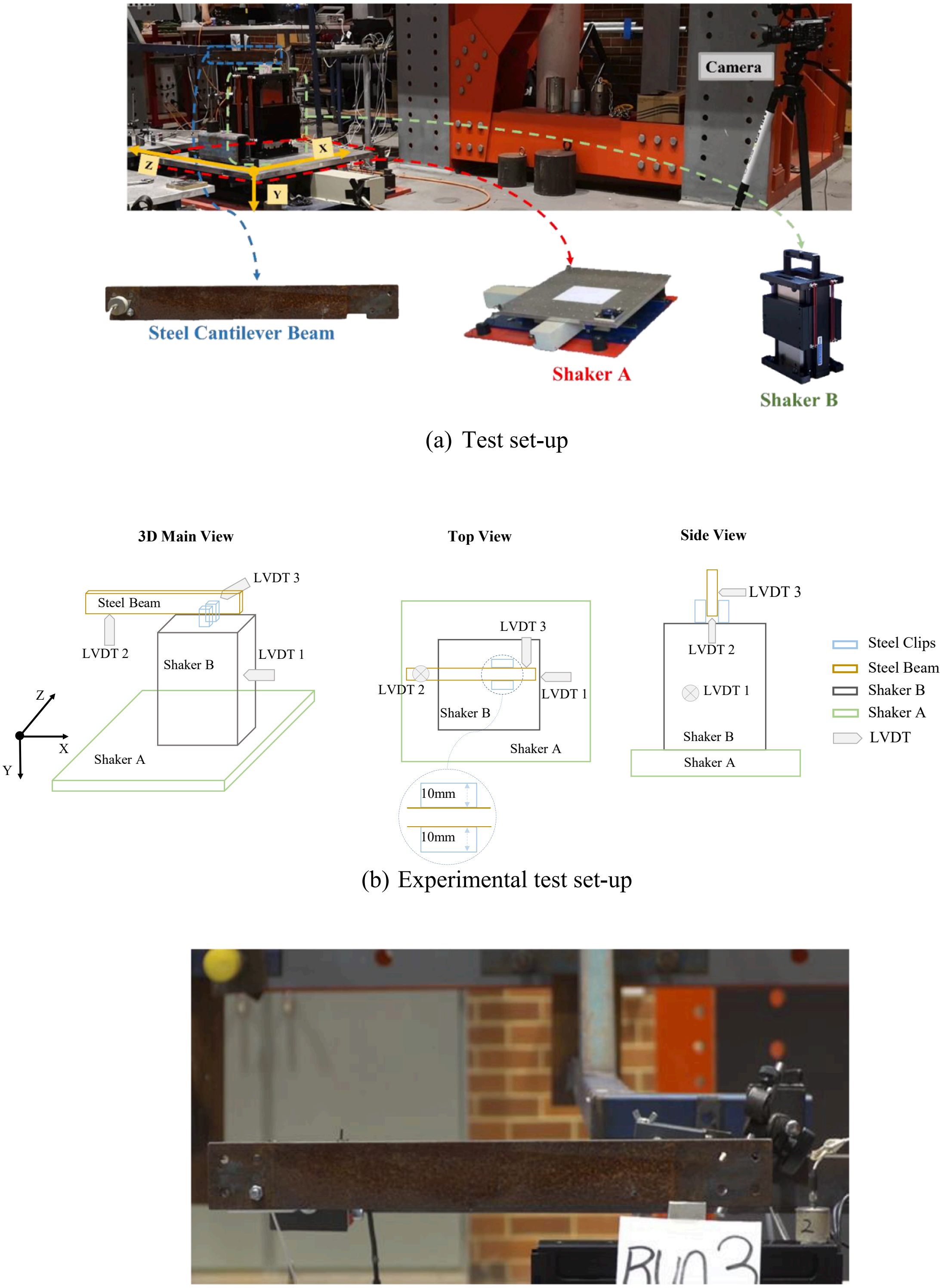
To estimate the ability of the proposed system for full-field measurement, 30 key points are detected using SuperPoint on the testing structure. It is hard to show the displacement trajectories of 30 key points in a figure. Instead, the numerical evaluations including relative errors and cross-correlation coefficients of all 30 key points are presented in Tables 4 and 5. Figures 12, 13 and 14 show the typical 3D displacement of Key point 15 (arbitrarily chosen) whose relative error and cross-correlation coefficient are given in the 15th row of Tables 4 and 5. The displacement responses measured by the corresponding physical sensors are used as the ground truth (G in the tables) with a relative error of 0.00% and a correlation coefficient of 1.0000. As shown in the tables, the proposed system is able to measure the in-plane displacement for all 30 key points in very high accuracy, with the average cross-correlation coefficients in the X direction as 0.9965 and Y direction as 0.9973, and the relative errors within 6%. For the out-of-plane displacement measurement, the average cross-correlation of the out-of-plane displacement measurement module is 0.8817, and in ten key points this metric achieves over 0.9. The average relative error of the out-of-plane displacement measurement is 50.09% against the physical sensor. These results demonstrate that using a single camera the proposed 3D vibration displacement measurement system can give very accurate measurements of in-plane vibration displacements. The accuracy of the measured out-of-plane vibration displacement is not as good as the in-plane displacement. The errors are mainly associated with the displacement amplitudes while the out-of-plane vibration frequencies are accurately measured. These results demonstrate the great potentials of the vision-based vibration measurement with only a single camera in monitoring the structural responses during extreme events, as well as under normal operation conditions. The measured data can be used in assessing the structural conditions.

Table 2
Cross-correlation coefficients and relative errors of ten key points.

No.	Corr. (ρ)	Relative Error. (%)	No.	Corr. (ρ)	Relative Error. (%)
G	1.0000	0.00	7	0.8287	57.04
1	0.8189	58.95	8	0.8269	53.27
2	0.8198	57.53	9	0.8248	52.10
3	0.8049	60.37	10	0.8287	49.77
4	0.8020	61.46	Ave	0.82003	55.41
5	0.8204	54.12	–	–	-
6	0.8252	49.41	–	–	-

Table 3
The version of the physical displacement sensors.

Sensor Name	Version	Measurement Direction
LVDT 1	HBM Displacement Transducer	X
LVDT 2	HBM Displacement Transducer	Y
LVDT 3	HBM Displacement Transducer	Z

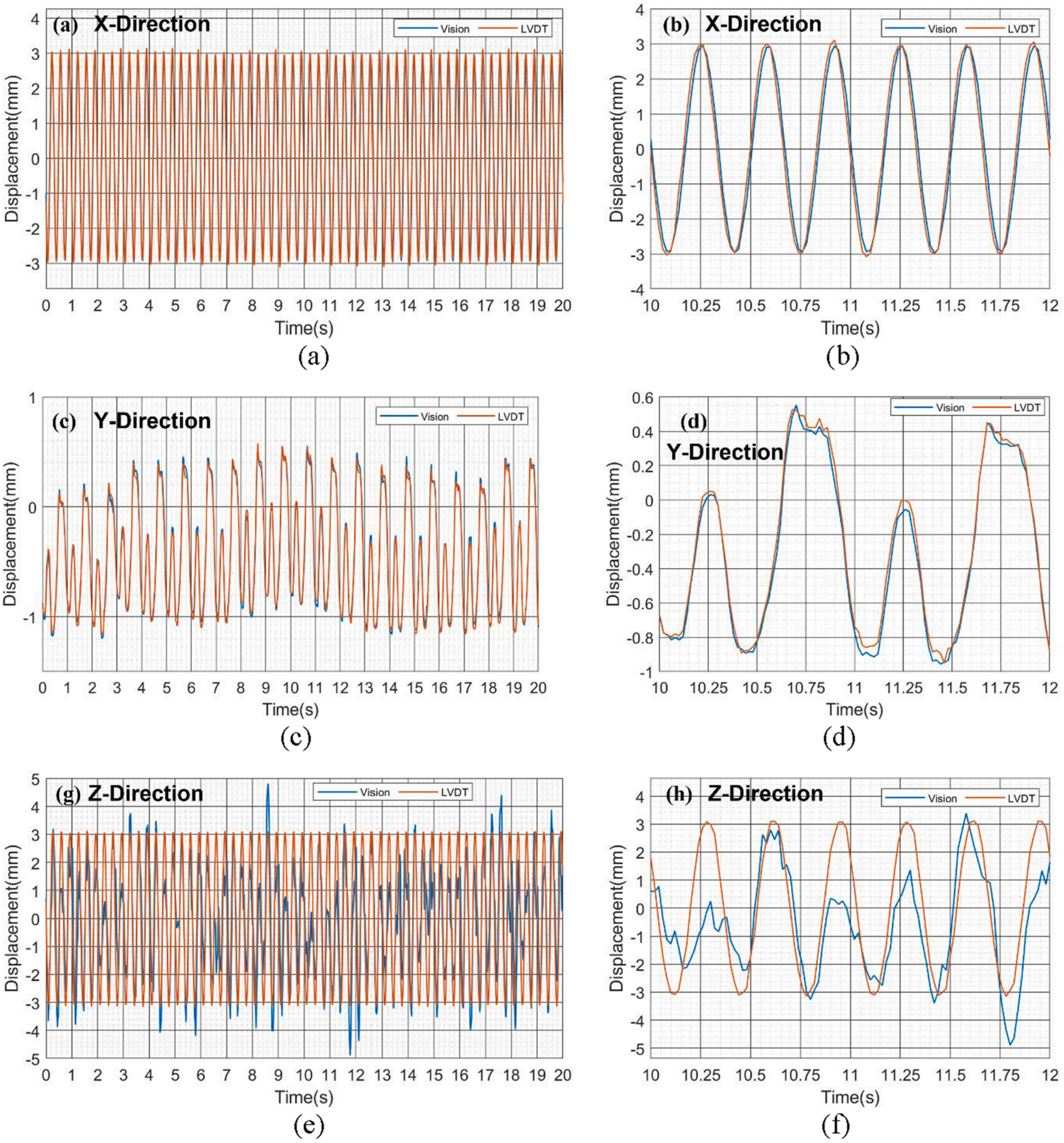


(a) Test set-up

(b) Experimental test set-up

(c) A frame of the camera's view

Fig. 11. Setup of Experimental test B.

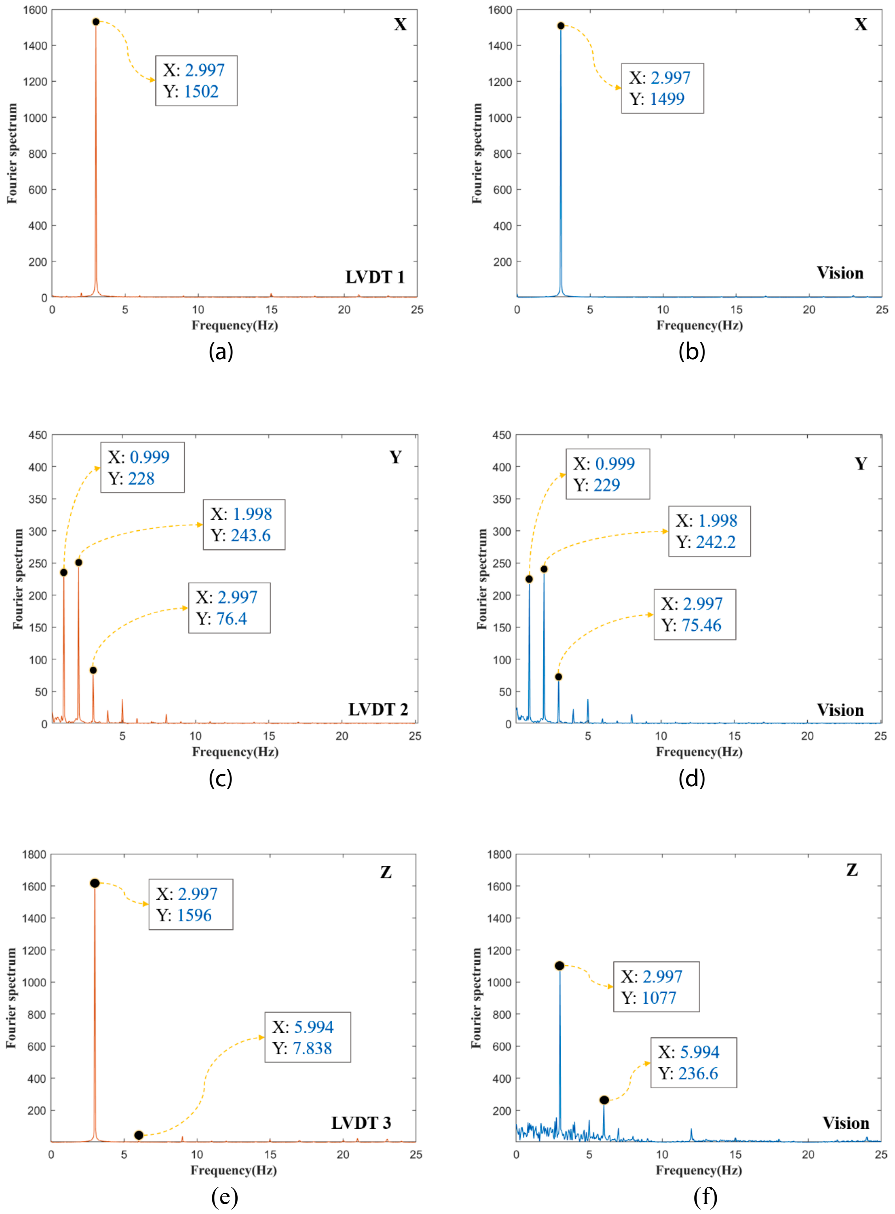


(a) (b)

(c) (d)

(e) (f)

Fig. 12. Evaluation of 3D displacement measurements at an arbitrary key point using the proposed vision-based system against physical displacement sensors: (a) Vision vs LVDT 1 in X direction; (b) A zoomed-in view of (a); (c) Vision vs LVDT 2 in Y direction; (d) A zoomed-in view of (c); (e) Vision vs LVDT 3 in Z direction; (f) A zoomed-in view of (e).



(a)

(b)

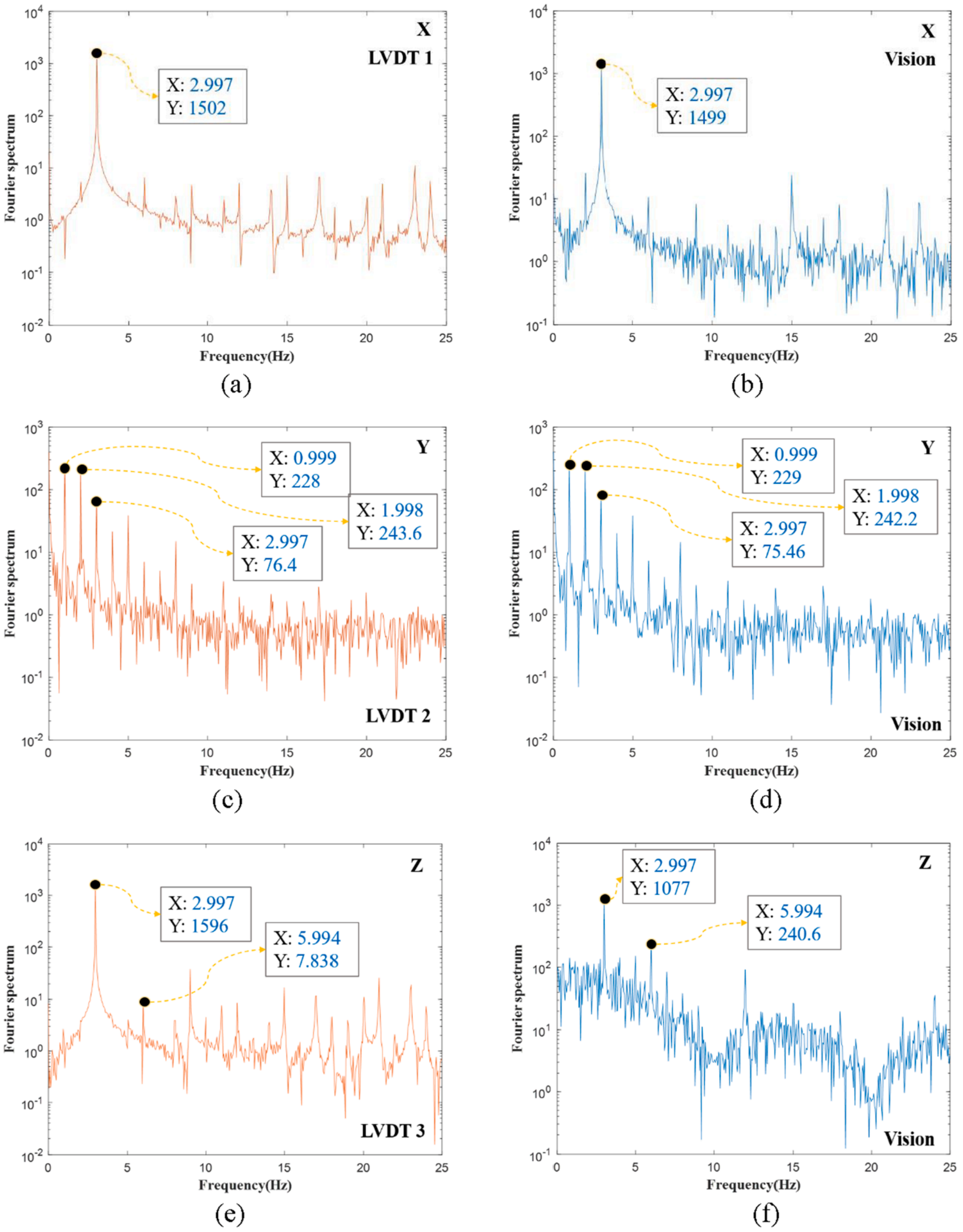
(c)

(d)

(e)

(f)

Fig. 13. Vibration frequencies obtained from the 3D vibration displacement measurement: (a) X direction by LVDT 1; (b) X direction by the proposed system; (c) Y direction by LVDT 2; (d) Y direction by the proposed system; (e) Z direction by LVDT 3; (f) Z direction by the proposed system.



(a) (b) (c) (d) (e) (f)

Fig. 14. Frequency spectrum in the logarithmic scale in 3D: (a) X direction by LVDT 1; (b) X direction by the proposed system; (c) Y direction by LVDT 2; (d) Y direction by the proposed system; (e) Z direction by LVDT 3; (f) Z direction by the proposed system.

Table 4
Cross-correlation coefficients of 30 key points in X, Y and Z direction.

No.	**Corr. (ρ)**			**No.**	**Corr. (ρ)**			**No.**	**Corr. (ρ)**		
	X	Y	Z		X	Y	Z		X	Y	Z
G	1.0000	1.0000	1.0000	**11**	0.9973	0.9974	0.8667	**22**	0.9984	0.9976	0.8528
1	0.9934	0.9973	0.9084	**12**	0.9965	0.9975	0.8858	**23**	0.9967	0.9984	0.8784
2	0.9983	0.9979	0.8969	**13**	0.9963	0.9979	0.9001	**24**	0.9974	0.9974	0.9097
3	0.9934	0.9982	0.9079	**14**	0.9984	0.9969	0.8834	**25**	0.9963	0.9983	0.8952
4	0.9994	0.9938	0.9099	**15**	**0.9975**	**0.9975**	**0.8850**	**26**	0.9974	0.9985	0.8300
5	0.9903	0.9962	0.8972	**16**	0.9944	0.9984	0.8844	**27**	0.9983	0.9975	0.8375
6	0.9938	0.9969	0.8890	**17**	0.9937	0.9934	0.8810	**28**	0.9973	0.9934	0.8058
7	0.9946	0.9973	0.9024	**18**	0.9982	0.9958	0.9064	**29**	0.9964	0.9985	0.8189
8	0.9983	0.9989	0.9051	**19**	0.9948	0.9984	0.8755	**30**	0.9988	0.9984	0.9024
19	0.9974	0.9973	0.8787	**20**	0.9987	0.9976	0.9036	**Ave**	**0.9965**	**0.9973**	**0.8817**
10	0.9984	0.9973	0.8918	**21**	0.9957	0.9985	0.8620	**–**	–	–	–

Table 5
Relative errors of 30 key points in X, Y and Z direction.

No.	**Relative Error (%)**			**No.**	**Relative Error (%)**			**No.**	**Relative Error (%)**		
	X	**Y**	**Z**		**X**	**Y**	**Z**		**X**	**Y**	**Z**
G	0.00	0.00	0.00	**11**	5.28	4.78	50.02	**22**	5.84	4.73	52.37
1	4.45	5.75	50.15	**12**	4.67	4.73	48.20	**23**	6.67	5.80	50.14
2	5.34	4.73	52.46	**13**	4.24	4.02	48.07	**24**	5.74	5.79	50.32
3	4.35	5.38	49.78	**14**	5.83	5.61	51.06	**25**	4.63	5.84	51.47
4	4.93	5.93	49.51	**15**	**5.05**	**5.79**	**49.85**	**26**	6.74	4.83	50.41
5	4.09	6.04	52.12	**16**	4.54	4.84	52.17	**27**	4.83	4.77	54.59
6	5.35	6.38	49.26	**17**	5.96	6.33	48.82	**28**	4.93	5.43	47.38
7	4.74	5.38	51.13	**18**	5.06	5.53	49.64	**29**	5.45	4.83	50.68
8	4.32	4.89	50.35	**19**	4.91	6.86	50.55	**30**	4.98	4.82	49.34
9	5.62	5.43	48.32	**20**	5.01	3.76	51.95	**Ave**	**5.13**	**5.36**	**50.09**
10	5.78	5.24	47.39	**21**	4.63	6.85	45.34	**–**	–	–	–

4. Conclusions

This paper proposes a monocular vision based 3D full-field displacement measurement system for civil structures. The in-plane displacement is measured using the advanced key point detection and tracking algorithms, while the advanced deep learning based monocular depth estimation technique is applied to measure the out-of-plane displacement utilising a single stationary camera. Experimental results demonstrate that the proposed monocular vision system can accurately measure the in-plane vibration displacement responses. The accuracy of the measured out-of-plane vibration displacement is not as good as the in-plane displacement, but the correlations between the ground truth data are all above 0.8 over the measurement duration of 20 s. The main contributions are as below:

1) The potential of data-driven methods for 3D displacement measurement based on monocular vision is demonstrated. Our system offers advantages over traditional displacement sensors and multi-view geometry-based methods in terms of accessibility and cost-effectiveness. This is particularly beneficial in scenarios where sensor installation is challenging, and budget constraints exist. It also avoids the complicated synchronisation requirement of using multiple cameras.
2) The second contribution of the proposed method lies in its depth independence and the ability to convert the estimated depth to absolute depth in engineering units. The proposed method leverages deep learning techniques to estimate the depth of objects in images by extracting depth cues, such as texture gradients, relative sizes, occlusion patterns and so on. These cues are inherently independent on the distance between the camera and the object. Additionally, the proposed method focuses on estimating the affine-invariant depth rather than the absolute depth from the camera to the structure. This ensures that the network is not specialized for a specific depth range, resulting in robust depth estimation across a wide range of distances.

The accuracy in the depth direction remains a primary area for improvement. The complexity of achieving accuracy in the depth direction arises from various factors, including limited training data and the absence of prominent depth cues in the scene. We plan to focus our future attention on the following two aspects:

1) It's essential to expand the dataset related to civil structures or scenes of civil structures. Gathering data from diverse structural environments, various lighting conditions, and different types of civil structures will enable the measurement system to better adapt to real-world scenarios. This comprehensive dataset can cover a wide range of structural variations, allowing the system to learn and adjust to different depth characteristics, therefore enable more accurate measurement of structural displacement in 3D.
2) The depth estimation neural network depends on the depth cues in the input image. Thoughtful camera positioning can enhance the presence of these depth cues in the scene, ultimately supporting the neural network in accurately estimating depth. However, summarizing and explaining the depth feature preference of complex neural networks can be a challenge. Neural networks operate in high-dimensional spaces and learn complex representations through multiple layers, making it hard to interpret the specific depth features learned by the network. Analytical and quantitative study of different depth feature preferences and the optimal setup of the system will allow for a thorough exploration of the inner workings of the depth estimation neural network and provide insights on how to enhance its performance.

CRediT authorship contribution statement

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

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] Feng D, Feng MQ. Computer vision for SHM of civil infrastructure: From dynamic response measurement to damage detection–A review. Eng Struct 2018;156:105–17.

[2] 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;190:210–22.

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

[4] Feng D, Feng MQ. Vision-based multipoint displacement measurement for structural health monitoring. Struct Control Health Monit 2016;23(5):876–90.

[5] Bartilson DT, Wieghaus KT, Hurlebaus S. Target-less computer vision for traffic signal structure vibration studies. Mech Syst Sig Process 2015;60:571–82.

[6] Tian Y, Zhang C, Jiang S, Zhang J, Duan W. Noncontact cable force estimation with unmanned aerial vehicle and computer vision. Comput Aided Civ Inf Eng 2021;36 (1):73–88.

[7] Park SW, Park HS, Kim JH, Adeli H. 3D displacement measurement model for health monitoring of structures using a motion capture system. Measurement 2015; 59:352–62.

[8] Yoon H, Elanwar H, Choi H, Golparvar-Fard M, Spencer Jr BF. Target-free approach for vision-based structural system identification using consumer-grade cameras. Struct Control Health Monit 2016;23(12):1405–16.

[9] Lv J, Lv M, Xiao J, Wen L, Lou Q. A point tracking method of TDDM for vibration measurement and large-scale rotational motion tracking. Measurement 2022;193: 110827.

[10] Narazaki Y, Gomez F, Hoskere V, Smith MD, Spencer BF. Efficient development of vision-based dense three-dimensional displacement measurement algorithms using physics-based graphics models. Struct Health Monit 2021;20(4):1841–63.

[11] Khuc T, Catbas FN. Computer vision-based displacement and vibration monitoring without using physical target on structures. Struct Infrastruct Eng 2017;13(4): 505–16.

[12] Ji YF, Chang CC. Nontarget image-based technique for small cable vibration measurement. J Bridg Eng 2008;13(1):34–42.

[13] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vis 2004;60(2):91–110.

[14] Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF). Comput Vis Image Underst 2008;110(3):346–59.

[15] Alcantarilla PF, Bartoli A, Davison AJ. KAZE features. In: European conference on computer vision. Berlin, Heidelberg: Springer; 2012. p. 214–27.

[16] DeTone D, Malisiewicz T, Rabinovich A. Superpoint: Self-supervised interest point detection and description. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops; 2018. p. 224–36.

[17] Hartley RI, Sturm P. Triangulation. Comput Vis Image Understand 1997;68(2): 146–57.

[18] Zhang Z. A flexible new technique for camera calibration. IEEE Trans Pattern Anal Mach Intell 2000;22(11):1330–4.

[19] Sarlin PE, DeTone D, Malisiewicz T, Rabinovich A. Superglue: Learning feature matching with graph neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition; 2020. p. 4938–47.

[20] Tsai YM, Chang YL, Chen LG. Block-based vanishing line and vanishing point detection for 3D scene reconstruction. In: 2006 international symposium on intelligent signal processing and communications. IEEE; 2005. p. 586–9.

[21] Tang C, Hou C, Song Z. Depth recovery and refinement from a single image using defocus cues. J Mod Opt 2015;62(6):441–8.

[22] Eigen D, Puhrsch C, Fergus R. Depth map prediction from a single image using a multi-scale deep network. Adv Neural Informat Process Syst 2014;27.

[23] Fu H, Gong M, Wang C, Batmanghelich K, Tao D. Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 2002–11.

[24] Liu F, Shen C, Lin G. Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2015. p. 5162–70.

[25] Chen W, Fu Z, Yang D, Deng J. Single-image depth perception in the wild. Adv Neural Informat Process Syst 2016;29:9.

[26] Xian K, Shen C, Cao Z, Lu H, Xiao Y, Li R, et al. Monocular relative depth perception with web stereo data supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. p. 311–20.

[27] Zoran D, Isola P, Krishnan D, Freeman WT. Learning ordinal relationships for mid-level vision. In: Proceedings of the IEEE international conference on computer vision; 2015. p. 388–96.

[28] Yin W, Wang X, Shen C, Liu Y, Tian Z, Xu S, … & Renyin D. Diversedepth: Affine-invariant depth prediction using diverse data. arXiv preprint arXiv:2002.00569. 2020.

[29] Yin W, Zhang J, Wang O, Niklaus S, Mai L, Chen S, et al. Learning to recover 3d scene shape from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 204–13.

[30] Yin W, Liu Y, Shen C, Yan Y. Enforcing geometric constraints of virtual normal for depth prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019. p. 5684–93.

[31] Lee S, Lee J, Kim B, Yi E, Kim J. Patch-wise attention network for monocular depth estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2021. p. 1873–81. No. 3.

[32] Yoneda K, Tehrani H, Ogawa T, Hukuyama N, Mita S. Lidar scan feature for localization with highly precise 3-D map. In: 2014 IEEE Intelligent Vehicles Symposium Proceedings. IEEE; 2014. p. 1345–50.

[33] Geiger A, Lenz P, Stiller C, Urtasun R. Vision meets robotics: The kitti dataset. Int J Robot Res 2013;32(11):1231–7.

[34] Silberman N, Hoiem D, Kohli P, Fergus R. Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. Berlin, Heidelberg: Springer; 2012. p. 746–60.

[35] Saxena A, Sun M, Ng AY. Make3d: Learning 3d scene structure from a single still image. IEEE Trans Pattern Anal Mach Intell 2008;31(5):824–40.

[36] Wang C, Lucey S, Perazzi F, Wang O. Web stereo video supervision for depth prediction from dynamic scenes. In: 2019 International Conference on 3D Vision (3DV). IEEE; 2019. p. 348–57.

[37] Hua Y, Kohli P, Uplavikar P, Ravi A, Gunaseelan S, Orozco J, & Li E. Holopix50k: A large-scale in-the-wild stereo image dataset. arXiv preprint arXiv:2003.11172. 2020.

[38] Ranftl R, Lasinger K, Hafner D, Schindler K, Koltun V. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Trans Pattern Anal Mach Intell 2020.

[39] Ilg E, Mayer N, Saikia T, Keuper M, Dosovitskiy A, & Brox T. Flownet 2.0: Evolution of optical flow estimation with deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017 pp. 2462–2470.

[40] Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision, Vol. 81, pp. 674–679.

[41] Tomasi C, Kanade T. Detection and tracking of point. Int J Comput Vis 1991;9: 137–54.

[42] Shi J. Good features to track. In 1994 Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. IEEE; 1994. pp. 593–600.

[43] 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; 538:117244.

[44] Liu Y, Zhuang B, Shen C, Chen H, & Yin W. Training compact neural networks via auxiliary overparameterization. arXiv preprint arXiv:1909.02214, 1. 2019.

[45] He K, Zhang X, Ren S, & Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. pp. 770–778.

[46] Deng Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009. pp. 248-255.

[47] Lin, G., Milan, A., Shen, C., & Reid, I. 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.

[48] Zamir AR, Sax A, Shen W, Guibas LJ, Malik J, & Savarese S. Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. pp. 3712–3722.

[49] Niklaus S, Mai L, Yang J, Liu F. 3d ken burns effect from a single image. ACM Trans Graph (ToG) 2019;38(6):1–15.

[50] Kim Y, Jung H, Min D, Sohn K. Deep monocular depth estimation via integration of global and local predictions. IEEE Trans Image Process 2018;27(8):4131–44.

[51] Xian K, Zhang J, Wang O, Mai L, Lin Z, & Cao Z. Structure-guided ranking loss for single image depth prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020. pp. 611–620.

[52] Li Z, & Snavely N. Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. pp. 2041–2050.

[53] Liu Z, Tang H, Lin Y, Han S. Point-voxel cnn for efficient 3d deep learning. Adv Neural Inf Proces Syst 2019;32.

[54] Dai A, Chang AX, Savva M, Halber M, Funkhouser T, & Nießner M. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. pp. 5828–5839.

[55] Simonyan K, & Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.

[56] Shi W, Caballero J, Huszár F, Totz J, Aitken AP, Bishop R … & Wang Z. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 1874–1883.

[57] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: European conference on computer vision. Cham: Springer; 2014. p. 740–55.

[58] Kingma DP, & Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. 2014.

[59] Kalal Z, Mikolajczyk K, & Matas J. Forward-backward error: Automatic detection of tracking failures. In: 2010 20th International Conference on Pattern Recognition, 2010. pp. 2756–2759.