High-Rate Displacement Monitoring with Low-Cost Multi-Vision Cameras using Global-Local Deep Deblurring and Rauch-Tung-Striebel Smoother

Peng "Patrick" Sun [1+], Mohammad Vasef [2+], Lin Chen [3]

[1] Assistant Professor, Department of Civil, Environmental, and Construction Engineering, University of Central Florida, 12800 Pegasus Drive, Suite 211, Orlando, FL 32816, email: peng.sun@ucf.edu, (corresponding author)

[2] Research Assistant, Department of Civil, Environmental, and Construction Engineering, University of Central Florida, 12800 Pegasus Drive, Suite 211, Orlando, FL 32816

[3] Associate Professor, Department of Bridge Engineering, Tongji University, Shanghai 200092, China

[+] co-first authors

Abstract:

Measuring structural vibrations help assess dynamic performances of civil structures and infrastructure. Although conventional displacement sensors have been widely adopted, they are contact-based methods which lack scalability. Recently, computer vision (CV) has been applied as a noncontact method to measure displacements. However, fast speed of structural vibration (e.g., in shake table tests) can inevitably cause motion blur that imposes challenges in all image-based object/feature detections, especially for normal portable cameras (without high-speed shutters). To address such issue, the study proposed a multi-vision, full-field sensing framework with affordable cameras using a novel global-local detection and deblurring (GLDD) module, which was designed with a generative adversarial network (GAN)-based deblurring model to enhance detection efficiency and accuracy by restoring blemished videos from multiple perspectives. Rauch-Tung-Striebel (RTS) smoother was studied for data fitting using incomplete observations caused due to severe motion-induced blurs. A shake table test was conducted on an aluminum frame with cameras and conventional sensors monitoring the structural vibrations. Fiducial markers were used to track the movement of the key locations on the structure. Results showed that the proposed method is satisfactory to monitor shake table tests when compared to conventional measurements with root-mean-square errors of 0.51-0.95 mm. The proposed deblurring module restored misdetection by 92.1%, 50.6%, and 25.2% for mild-, medium-, and severe-level motion blurs, respectively. Smoother-based data fitting outperformed filter-based one when dealing with highly blemished images. The proposed monitoring system with GLDD and RTS smoother-based data fitting provides a robust measurement solution when dealing with motion blurs.

Keywords:

shake table test, motion blur, computer vision, generative adversarial networks, structural health monitoring.

Highlights:

- proposed a multi-vision displacement measurement approach with global-local detection and deblurring (GLDD) module using GAN-based deep deblurring method to address the motion blur issue
- developed an automated algorithm for affordable cameras to monitor displacement in shake table tests, including feature detection, global-local image deblurring, multi-video synchronization, and filter/smoother-based data fitting
- studied the performances of different data fitting methods on displacement measurements for severe motion blur cases using Kalman filter and Rauch-Tung-Striebel (RTS) smoother
- provided the guidelines for using the proposed approach and affordable cameras to achieve displacement monitoring in shake table tests

1. Introduction

Monitoring structural responses (e.g., displacement, acceleration, strain) is used to assess the behavior of civil structures. Measured data from experimental tests (e.g., quasistatic test, shake table test) are usually influenced by the characteristics and limitations of the adopted measurement methods (Zona, 2020). Structural responses are commonly measured using wired, contact sensors at desired locations of a structure. Non-contact measurement methods take one step further by avoiding the physical contact between sensor and structures, such as strain sensors using computer vision (CV) techniques (e.g., digital image correlation (del Rey Castillo et al., 2019), photoluminescence techniques (Sun et al., 2019), and laser Doppler effect (Xu et al., 2019). In addition, existing displacement measurement methods include linear variable differential transformer (LVDT), real-time kinematic (RTK) global navigation satellite systems (GNSS) /global positioning system (GPS) sensors (Bezcioglu et al., 2023), terrestrial laser scanner (Kogut & Pilecka, 2020), and double-integration from acceleration (Zheng et al., 2019). However, these displacement measurement methods exhibit specific limitations, such as low-sampling rate (Ma et al., 2022) of RTK-GNSS, limited accuracy in GPS measurements (Rychlicki et al., 2020), high-noise level in terrestrial laser scanner (Muralikrishnan, 2021), (potential) large low-frequency drift using double integration of accelerations (Zheng et al., 2019), and deployment cost of laser Doppler-based method (Chu, 2005). In addition, accessibility issues (especially in long bridges and high-rise buildings), cost escalation for up-scale measurement, range constraint, and generally the requirement for a stable installation platform are the complexities to consider when utilizing LVDTs.

Vision-based methods were studied to obtain displacement measurement and overcome some of the limitations. In recent years, the technological progress in computing power, computer vision algorithms (Sun et al., 2022), and high-speed cameras (Zhang et al., 2016) attracted more attentions on the direct measurement methods (Greenbaum et al., 2016) and further applications on the measurements [e.g., system identification (Yang et al., 2019), finite element model updating (Dong et al., 2020), damage detection (Guo et al., 2019)] of vision-based methods. Vision-based applications in shake table tests started from early 2010's by adopting early-stage feature detection algorithms, (large-size) primitive artificial tags, and localization methods to measure structural displacements (Choi et al., 2011). Structural vibration of full-scale civil infrastructures or large scaled models are usually neither in high speed nor in high frequency [e.g., frequency range for most civil infrastructure is well below 70 Hz (Zona, 2020) or even much lower as several Hz]. Therefore, most of the time a portable camera with a low frequency capacity [e.g., 30 frame-per-second (fps)] is sufficient and the blur due to structural vibration will not be a serious issue for displacement monitoring. Most of the current studies focused on the further structural health monitoring (SHM) applications of CV-based displacement measurement (e.g., behavior analysis, load estimation, modal identification, model updating, damage detection) (Dong & Catbas, 2021) and much fewer studies focused on solving practical issues in monitoring applications in shake table tests, such as perspective selection (e.g., single-vision, dual-vision), camera location/pose limitation, illumination condition, occlusion, video frame asynchronization (if there is multiple cameras), and motion-induced image blur.

Some of these issues can be addressed in a controlled lab environment during a shake table test, for example, using proficient direct current (DC) lights to provide adequate illumination and using post-synchronization technique to solve asynchronization issue. However, some other challenges remain to be resolved. For example, experimental studies on structural dynamics in particular, can suffer from motion-induced image blurs. Because shake table tests are conducted in lab environments and researchers may use smaller-scale models subjected to more intense excitations especially when near resonance, making collected video data much more susceptible to the issue of motion-induced blur. Most of time researchers would adopt pricy, high-speed cameras (~$8-30k, 200-2000 fps) to avoid the motion blurs in their CV applications for shake table without solving the issue. However, even with the most advanced camera with high-speed shutter, the motion blur issue is still there when dealing with any fast-moving object relative to the camera shutter. The studies on image deblurring using post-processing techniques for motion-induced blurs are found to be very rare if there is any. Hence, a remedy solution is in great need to address the negative impacts from motion blurs for most current shake table users with affordable measurement setups, such as portable cameras with normal speed (~$1-2k, 30-60 fps).

The objective of this study was to develop a vision-based displacement monitoring framework for shake table tests that is robust to the motion blur issue. The study proposed a multi-vision approach with the ability to remediate motion-blur effect using deblurring module and data fitting module for accurate displacement monitoring. This paper firstly introduced a multi-vision sensing approach with a global-local detection and deblurring (GLDD) module to reduce the effect of motion blur and a data fitting module to estimate midsection based on incomplete observation. Secondly, the study designed a shake table test to evaluate the proposed sensing approach on an aluminum frame structure with different severity levels of vibration. Further, the discussion based on the augmented measurements and data fitting was conducted and useful guidelines was provided as well. In the end, the paper provided a summary for this work as well as its limitation and future work.

2. Problem Statement

(1) Motion-Induced Blur in Images

Typical digital image generation includes two steps: image signal acquisition and color rendering with image signal processor (ISP). In a classical pin-hole camera model (Ma et al., 2004) (Fig. 2a), a 3D point in the World Coordinate System (WCS) is denoted as P and the point is visualized as a pixel (p') projected onto the camera sensor plane in a 2D Sensor Coordinate System (SCS). The relationship of P and p' can be expressed in a concise form as:

$$ {}^{s}\mathbf{x}_{p'} \sim \mathbf{M}_{\mathrm{aff}}\, \mathbf{M}_{\mathrm{proj}} [\mathbf{R}|\mathbf{T}]\, {}^{w}\mathbf{X}_{P} \quad (1) $$

where $\mathrm{M_{aff}}$ and $\mathrm{M_{proj}}$ are affine matrix and projection matrix containing intrinsic parameters, $[\mathbf{R}|\mathbf{T}]$ is the joint rotation-translation matrix (containing extrinsic parameters), ${}^{s}\mathbf{x}_{p'} = \left[{}^{s}x_{p'}, {}^{s}y_{p'}, 1\right]^T$ and ${}^{w}\mathbf{X}_P = [{}^{w}X_P, {}^{w}Y_P, {}^{w}Z_P, 1]^T$ denote the corresponding homogeneous coordinates, respectively.

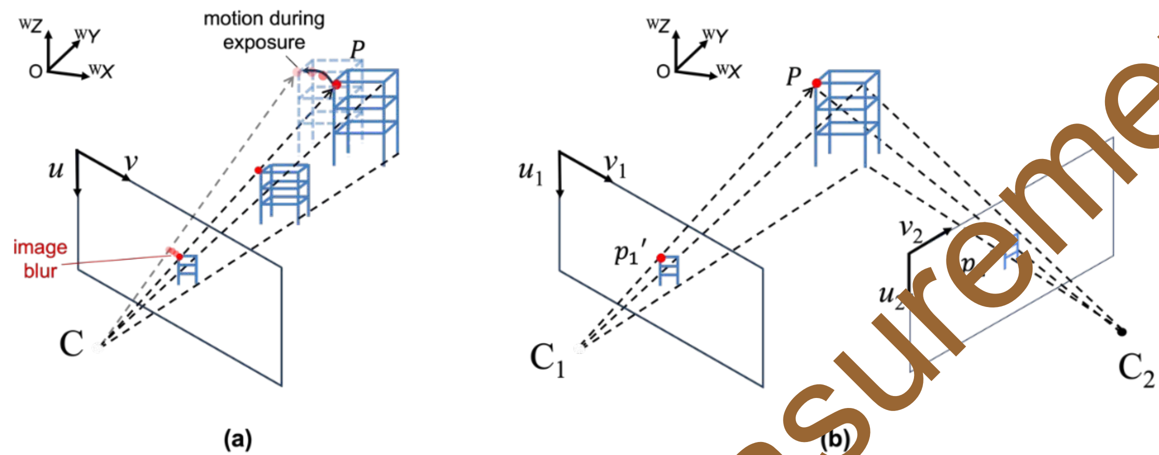


Figure 1. Schematic views of (a) single-perspective setting with motion-induced blur on the sensor coordinate system (SCS), and (b) dual-perspective setting with the same structure feature observed by both cameras.

Image blurs can result from the relative movement between camera and object/scene and it can be formulated as the accumulation of photons on camera sensors during the exposure time:

$$I_b(x, y) = \mathrm{ISP}\left(\int_{t_1}^{t_2} f(t, x, y)\, dt\right) \tag{2}$$

where I_b denotes the blurred image, $[t_1, t_2]$ represents the time window for exposure, $f(t, x, y)$ represents the photon response at pixel location (x, y) at time instant t, and $\mathrm{ISP}(\cdot)$ denotes the image signal processor operator (e.g., white balance, color correction).

Motion-induced image blurring (denoted in **Figure 1**a) is influenced by the shutter speeds of cameras and the relative movement speeds between cameras and recorded objects. Motion blurs can be categorized into local blurs and global blurs. Global blurs usually occurs with a moving camera that is usually encountered in the application of robotic vision (Zeng et al., 2020) and simultaneous localization and mapping (SLAM) (Gao & Zhang, 2021). While local blurring results from moving objects in static backgrounds. Local blur issue occurs in the CV application for shake table tests where the table base and the mounted dynamic structures are the foreground in motion and cameras are fixed statically with the background.

Artificial features, such as fiducial markers, have strong (black-white) contrast and sharp features with straight edges for robust detection compared to natural features. However, severe structural motion during a shake table test can make fiducial marker detection in blurred images a really challenging task. Motion blur can deteriorate the marker detection performance using image processing algorithms (e.g., edge detection, blob detection). Deblurring methods can render clearer images for actuate feature detection. The underlying problem of the image deblurring part in this study is to restore clearer and sharper visual features on images for accurate detection and displacement computation.

(2) Motion Estimation for Severe Blur

In practice, severe motion blur on images can cause misdetections even when deblurring technique is implemented. For example, when the structure is subjected to vibration with a frequency close to the natural frequencies, structural vibration becomes much severer making the top floor shaking faster than the other floors. A post processing of data fitting is needed to complement the measurement in these cases. Although linear interpolation and/or spline interpolation can be used as basic data fitting considering the continuous movement of structure using nearby measurements, these interpolation methods neglect the system information and sometimes can yield wrong estimates at the misdetection instances.

Assume a measurement from a shake table test is denoted as $\mathbf{y}_k$ at instance $t_k (k = 1,2,\dots,T)$ and the corresponding state is denoted as $\mathbf{x}_k$. For example, the measurement includes displacement measurement in x direction $\mathbf{y}_k = d_x(t_k)$ for a 1D shake table test and the state include both displacement and velocity, such as $\mathbf{x}_k = [d_x(t_k), \dot{d}_x(t_k)]$. To model the motion, a state vector $\mathbf{x}_k \in \mathbb{R}^{n\times 1}$ denotes the system state and the linear dynamic system can be expressed as:

$$\mathbf{x}_k = \mathbf{A}_{k-1}\mathbf{x}_{k-1} + \mathbf{q}_{k-1} \tag{3}$$

where $\mathbf{x}_k \in \mathbb{R}^{n\times 1}$ is the state (as a vector), $\mathbf{q}_{k-1} \in \mathbb{R}^{n\times 1}$ is the process noise with Gaussian probability distribution $\mathbf{q}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{Q}_{k-1})$, and $\mathbf{A}_{k-1} \in \mathbb{R}^{n\times n}$ denotes the dynamic model/transition matrix.

The measurement equation is:

$$\mathbf{y}_k = \mathbf{H}_k\mathbf{x}_k + \mathbf{r}_{k-1} \tag{4}$$

where $\mathbf{y}_k \in \mathbb{R}^{m\times 1}$ is the measurement, $\mathbf{r}_k \in \mathbb{R}^{m\times 1}$ is the measurement noise with Gaussian probability distribution $\mathbf{r}_{k-1} \sim \mathrm{N}(\mathbf{0}, \mathbf{R}_{k-1})$, and $\mathbf{H}_k \in \mathbb{R}^{m\times m}$ denotes the measurement model/matrix.

Successful observations are denoted as $(\mathbf{y}_k)_i \equiv (\hat{\mathbf{y}}_k)_i, (k,i) \in \mathcal{K}$:

$$(\mathbf{y}_k)_i = (\mathbf{H}_k \mathbf{x}_k + \mathbf{r}_{k-1})_i \qquad (k,i) \in \mathcal{K} \tag{5}$$

where $\mathcal{U} = \{1,2,\dots,T\} \times \{1,2,\dots,m\}$ denotes the universal set of scalar outputs corresponding to all possible observations, $\mathcal{K} \subseteq \mathcal{U}$ denotes the set corresponding to successful observations, $\mathcal{M}$ denotes the set corresponding to failed/missed observations. $\mathcal{K} \cap \mathcal{M} = \emptyset$ and $\mathcal{K} \cup \mathcal{M} = \mathcal{U}$.

Severe structural vibrations sometimes can make motion blur so blemished that CV-based measurements cannot be obtained successfully even after using some image deblurring techniques. For discussion, these failed/missed observations are denoted as $(\mathbf{y}_k)_i \equiv (?_k)_i$, $(k,i) \in \mathcal{M}$. The underlying mathematical problem of the data fitting part in this study is to estimate failed/missed measurements $(?_k)_i, (k,i) \in \mathcal{M}$ based on the successful observations, $(\hat{\mathbf{y}}_k)_i, (k,i) \in \mathcal{K}$.

3. Method

Motion blurs involve shutter speed of camera and moving speed of recorded object. The failure in feature detection resulting from motion blur can cause misdetection of key features on vibrating structures, leading to empty observations at certain time instances. Hence, it is important to develop a deblurring and detection strategy that is suitable for displacement monitoring in shake table tests. Targeting toward the motion blur issue in shake table tests, a multi-vision approach was proposed including three modules (as shown in **Figure *2***): multi-vision photogrammetry module, global-local deblurring and detection module, and data fitting module for severe motion blurs.

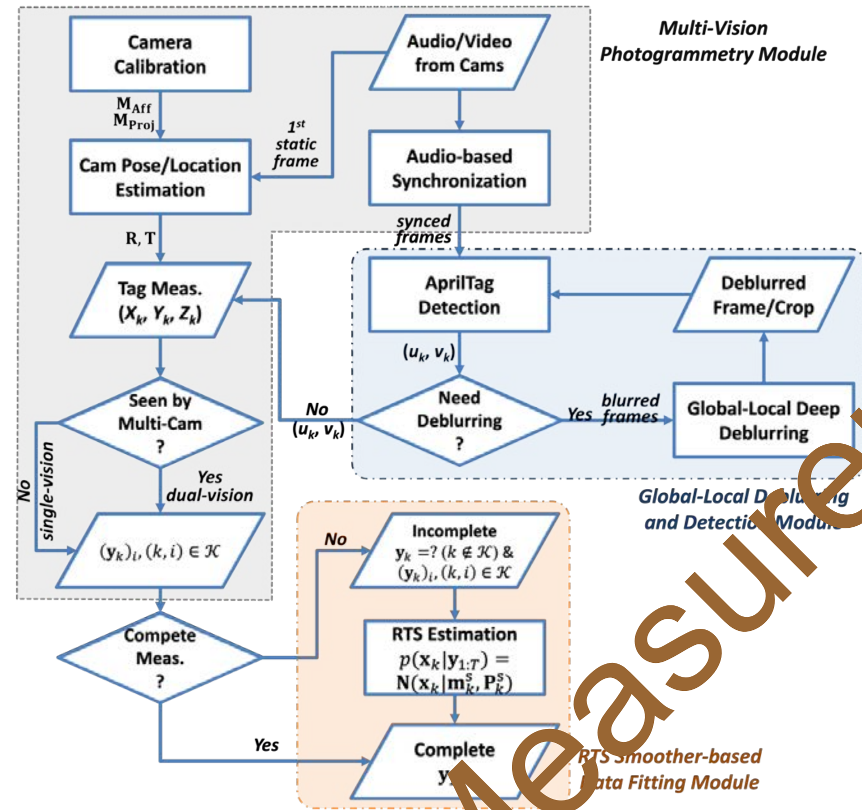


Figure 2. Flow chart of the proposed approach for displacement monitoring with muti-vision photogrammetry module, image deblurring and detection module, and data fitting module.

3.1. Multi-Vision Displacement Monitoring

In a shake table test, there are situations that not all tags are in the scope of view. For example, the top floor of a structure might be out of the camera due to excessive displacements, or one feature for tracking may be blocked by a structure component from one camera view. Sometimes during experiments, visual features on structures (e.g., artificial patterns, natural features) cannot be seen clearly or easily due to limited conditions (e.g., poor illumination, unsatisfactory camera pose, large movement of structure). To cope with these non-perfect situations in practice, a multi-vision strategy with both single-vision and dual-vision choices is needed to obtain full-field measurement in the post data analysis. In addition, vibrating structure may induce different levels of blurs viewed in multiple perspectives. Even if motion blur (**Figure *1***a) is too severe to be viewed clearly in one perspective, it doesn't necessarily mean that the blur will be at the same level with another perspective. The effectiveness of image deblurring of the same visual features may differ due to different perspectives. Therefore, multi-vision scheme was chosen to provide information redundancy for dynamic experiments.

(1) Fiducial Marker Detection and Video Synchronization

Visual features in images used for displacement measurements could either be natural features (e.g., structural corners) or artificial features (e.g., fiducial markers). Marker-free methods require no speckle pattern or marker deployment, but they require more computation time to process images to get features for matching. In contrast, artificial markers are commonly used to obtain streamlined detection and tracking of points of interest (Spencer Jr et al., 2019). In order to process videos in a fast manner, a fast feature detection and association (across cameras) algorithm is preferred for shake table tests. Therefore, fiducial markers (e.g., AprilTag (Olson, 2011)) with sharp features was on top of the list for this study, as well as a speedy detection algorithm. AprilTag detection algorithm (Wang & Olson, 2016) includes the first step of quad detection and the second step of detailed pattern recognition. In the second step, a quad candidate generated from the first step will be decoded to compare with the tag dictionary in the family to decide if the binary payload matches with one specific tag pattern. This work aims to address the issue of motion-induced blur in shake table tests and the proposed method can be integrated with different types of visual features for CV-based monitoring. For demonstration purpose, AprilTag's were used as example for feature detection.

In the first step of tag detection, quad detection may fail if line/quad features are blemished by motion blur. In the second step, even if a marker is detected as a candidate quad in the first step, the decoder will filter a marker out if its binary pattern is wrecked by motion blur, leading to no match in the known tag family (Krogius et al., 2019; Liu et al., 2022). Therefore, there is a need to restore images and recover the sharp features of the markers before achieving tag detection.

Followed by structural feature detection, video synchronization is of great importance for muti-vision application, especially for shake table tests. One may argue to have all the cameras are triggered at the same time in the beginning of the shake table tests to enforce video frames match. However, the internal clock within each of the cameras will slightly drift during the recording (especially for non-expensive cameras), making a mismatch of video frame across different cameras. The mismatched frames will yield considerable error in multi-vision triangulation computation. Therefore, a post video synchronization is needed before the image processing. In this study, the ambient sound from shake table tests was recorded on the audio channels and the audio recordings were processed and synchronized based on audio waveform matching using cross-correlation.

(2) Multi-Vision Triangulation

In a dual-vision setting, images from two different perspectives (**Figure** *1*b) can serve as strong constraint in 3D scene reconstruction when the two viewing rays corresponding to the same scene point intersect. The 3D coordinates can then be determined using the direct linear transform (DLT) method (Abdel-Aziz et al., 2015) based on triangulation.

$$ {}^s\mathbf{x}_{p'} = \begin{bmatrix} A^{T^k}\,{}^w\mathbf{X}_P \\ B^{T^k}\,{}^w\mathbf{X}_P \\ C^{T^k}\,{}^w\mathbf{X}_P \end{bmatrix} \tag{6} $$

where $A^{T^{(k)}}$, $B^{T^{(k)}}$ and $C^{T^{(k)}}$ represent the three rows of the transformation matrix $\mathbf{M}_{\text{trans}}{}^{(k)}$ for k-th camera. Transformation matrix $\mathbf{M}_{\text{trans}} = \mathbf{M}_{\text{aff}}\,\mathbf{M}_{\text{proj}}[\mathbf{R}|\mathbf{T}]$.

In this study, pinhole model (Eq. 1) was used for each of the two points in the two images (e.g., p_1', p_2' in **Figure** *1*b), respectively.

$$ \begin{cases} {}^s x_{p_i'} = \dfrac{{}^s u_{p_i'}}{{}^s w_{p_i'}} = \dfrac{A^{T^k}\,{}^w X_P}{C^{T^k}\,{}^w X_P} \\ {}^s y_{p_i'} = \dfrac{{}^s v_{p_i'}}{{}^s w_{p_i'}} = \dfrac{B^{T^k}\,{}^w X_P}{C^{T^k}\,{}^w X_P} \end{cases} \rightarrow \begin{cases} ({}^s x_{p_i'} C^{T^k} - A^{T^{(k)}})\,{}^w X_P = 0 \\ ({}^s y_{p_i'} C^{T^k} - B^{T^{(k)}})\,{}^w X_P = 0 \end{cases} \tag{7} $$

Combining the equations developed from the two points (Eq. 7), linear algebra equations are derived to yield a unique solution. The four observations (${}^s u_{p'}$ and ${}^s v_{p'}$ from each point) make it a determinate problem to solve.

$$ \begin{bmatrix} {}^s x_{p_1'} C^{T^{(1)}} - A^{T^{(1)}} \\ {}^s y_{p_1'} C^{T^{(1)}} - B^{T^{(1)}} \\ {}^s x_{p_2'} C^{T^{(2)}} - A^{T^{(2)}} \\ {}^s y_{p_2'} C^{T^{(2)}} - B^{T^{(2)}} \end{bmatrix}_{4\times 4} \begin{bmatrix} {}^w U_P \\ {}^w V_P \\ {}^w W_P \\ {}^w T_P \end{bmatrix}_{4\times 1} = \mathbf{0} \tag{8} $$

where ${}^w T_P$ is treated as a scale factor and the homogeneous coordinates of point P could be represented as $[{}^w X_P\,, {}^w Y_P\,, {}^w Z_P\,, 1]^T$.

3.2. Global-Local Detection and Deblurring (GLDD)

Restoring images from local blurring of moving objectives is an open problem and image-based deep deblurring methods are yet to be applied in CV-based monitoring of shake table tests. One focus of this study lies on the design of global-local detection and deblurring (GLDD) module using deep deblurring to augment the displacement measurement in shake table tests.

(1) Deep Deblurring Model

In the study, a deblurring model was adopted that is a generative adversarial network (GAN)-based deep deblurring model. GANs were investigated for image restoration (Ramakrishnan, 2017) by refereeing to the idea of image translation and the recent development includes DeblurGAN (Kupyn et al., 2018), DeblurGAN v2 (Kupyn, 2019), and Ghost-DeblurGAN (Liu et al., 2022). Due to the high speed of processing, models with light-weight convolution neural network (CNN)-based feature extractors, such as GhostDeblurGAN, are preferred in the study for efficient image deblurring compared with heavyweight models. Hence, this study adopted a lightweight deblurring model, DeblurGAN-v2, as the image restoration component for the proposed deblurring module.

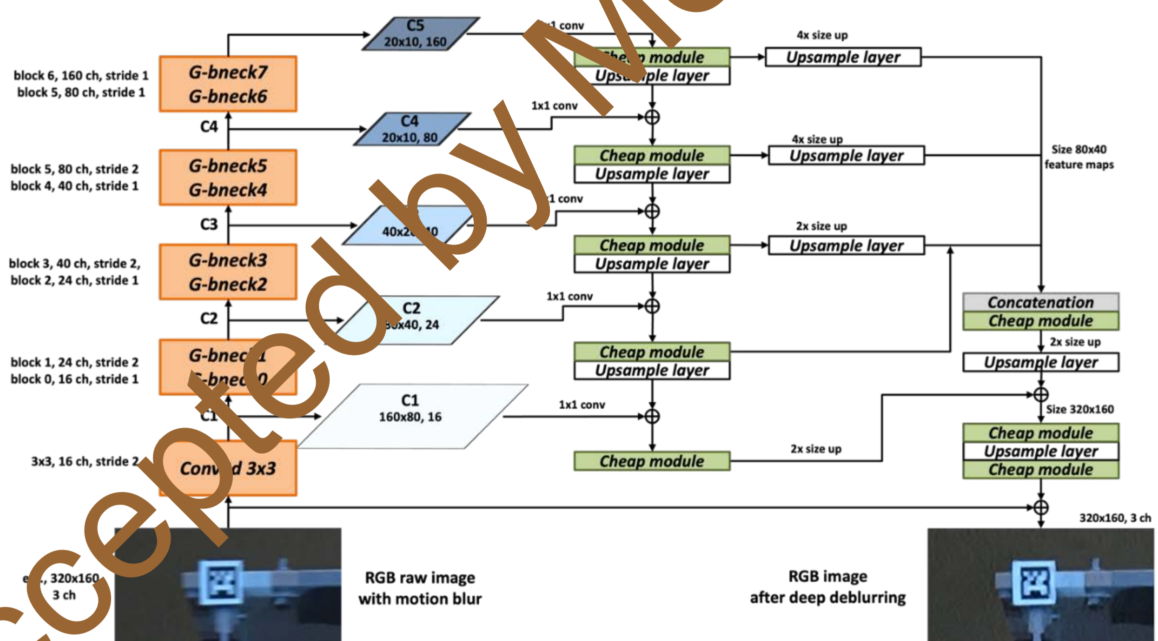


Figure 3. Schematic view of the Ghost-DeblurGAN generator architecture with an example of image processing on fiducial marker attached to a structure.

The generator part of a GAN-based deep deblurring model includes a CNN backbone as feature extractor and a feature pyramid network for a rich set of global feature maps at different spatial scales (**Figure 3**). Compared to DeblurGAN-v2, the CNN backbone of GhostDeblurGAN uses GhostNet (Han, 2020) as shown in **Figure 3** instead of MobileNet. Cheap module (Sandler, 2018) adopted in the GhostNet CNN backbone includes the pointwise and depthwise separable convolution layers in sequence to obtain the intrinsic feature maps. These intrinsic feature maps are comparable to computational expensive 2D convolution layers, but the yielding has a much faster speed. Compared to the 2D convolution layers in DeblurGAN-v2, the cheap modules in Ghost-DeblurGAN can help reduce 53.21% of the floating-point operations per second (FLOPS) during the forward calculation. Hence, Ghost-DeblurGAN model was adopted as the deblurring model of the proposed framework. The deblurring model was trained using a large-scale YorkTag dataset (Liu et al., 2022) with paired blurry-sharp tag images of 2074 pairs (1577 in training set and 497 in test set) that were collected in both indoor and outdoor environments. The deblurred images (9761 tags within dataset) processed using the trained deblurring model showed an improved detection rate of 59.1%, compared to detection rate of only 32.0% when using raw blurred images.

(2) Global-Local Detection and Deblurring (GLDD) Module

A global detection and deblurring (GDD) process on the whole image may be sufficient to restore images from small motion blurs. However, when the motion blur becomes severer, local level of deblurring process is needed on the key locations to retrieve sharp visual clue for tag detection. Therefore, global-level and local-level image deblurring were automated to augment tag detection based on the different extent of motion blur.

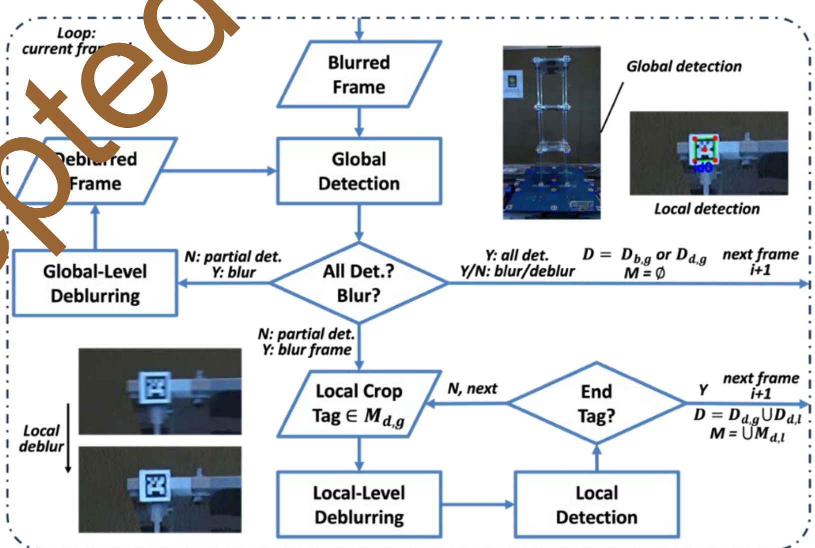


Figure 4. Flow chart of the proposed GLDD module. (note: both artificial and natural features can be used and AprilTag is adopted as an example)

The proposed GLDD module (**Figure *4***) includes global-level tag detection/deblurring on whole video frames and local-level detection/deblurring on cropped images near the key (image) locations. The assumption for such design is that the attention of a general deep deblurring model is distracted on nonimportant area when dealing with a relatively large moving foreground (e.g., the vibrating structure) instead of focusing on the key structural features. If an image crop contains single features/tags with relative larger foreground, the attention of a CNN-based feature extractor will be forced to put on the tags over the background and the tags will be less difficult to restore. For a shake table test, rigid movement occurs at the sliding base. If the structure moves at the same frequency as the sliding base when subjected to a forced vibration, the moving distance on the higher levels of the structure over the same time would be larger compared to the sliding base. Hence, it is reasonable to assume that the blur severity at structure top is larger than the base.

3.3. Data Fitting for Severe Motion Blur

The problem of data fitting was reshaped with a perspective of Kalman filtering (KF) (Welch & Bishop, 1995) and Rauch-Tung-Striebel (RTS) smoothing (Särkkä, 2008) thinking. In this study, the dynamic system was described by a partially observed Markov process in the Bayesian sense by computing the conditional distributions (e.g., $p(\mathbf{x}_k|\mathbf{x}_{k-1})$) using either filtering or smothering methods (Särkkä & Svensson, 2023).

(1) Kalman Filter-based Estimation

KF can estimate current state of a dynamical system ($\mathbf{x}_k$) given previous and current observations ($\mathbf{y}_j, j = 1,2,\dots,k$). KF probabilistic state model consists of the conditional probability distributions of the state and the measurement which are Gaussian distributions:

$$\mathbf{x}_k \sim p(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathbf{N}(\mathbf{x}_k|\mathbf{A}_{k-1}\mathbf{x}_{k-1}, \mathbf{Q}_{k-1}) \quad (9)$$

$$\mathbf{y}_k \sim p(\mathbf{y}_k|\mathbf{x}_k) = \mathbf{N}(\mathbf{x}_k|\mathbf{H}_k\mathbf{x}_k, \mathbf{R}_k) \quad (10)$$

In order to compute the filtering results, the parameters or states are computed in two steps recursively: prediction step and correction step. The prediction step in the recursive computation includes the mean prediction and covariance prediction:

$$\mathbf{m}_k^- = \mathbf{A}_{k-1}\mathbf{m}_{k-1} \tag{11}$$

$$\mathbf{P}_k^- = \mathbf{A}_{k-1}\mathbf{P}_{k-1}^-\mathbf{A}_{k-1}^T + \mathbf{Q}_{k-1} \tag{12}$$

The difference between the predicted measurement and the sensor reading is denoted as innovation $\mathbf{v}_k$:

$$\mathbf{v}_k = \mathbf{y}_k - \mathbf{H}_k\mathbf{m}_k^- \tag{13}$$

The correction step in the recursive computation includes the correction of mean and covariance of the current state:

$$\mathbf{m}_k = \mathbf{m}_k^- + \mathbf{K}_k\mathbf{v}_k \tag{14}$$

$$\mathbf{P}_k = \mathbf{P}_k^- - \mathbf{K}_k\mathbf{S}_k\mathbf{K}_k^T \tag{15}$$

where $\mathbf{m}_k^-, \mathbf{m}_k \in \mathbb{R}^{n\times 1}$ are the mean value of the state $\mathbf{x}_k$ and $\mathbf{P}_k^-, \mathbf{P}_k \in \mathbb{R}^{n\times n}$ are the covariance matrix of the measurement $\mathbf{x}_k$ during the prediction and correction steps, respectively. Kalman gain for the correction is $\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}_k^T\mathbf{S}_k^{-1}$. Covariance matrix for the innovation is $\mathbf{S}_k = \mathbf{H}_k\mathbf{P}_k^-\mathbf{H}_k^T + \mathbf{R}_k$.

From the previous state $\mathbf{x}_{k-1}$ and current measurement $\mathbf{y}_k$, a prediction and correction can be performed to estimate current state of $\mathbf{m}_k$ and $\mathbf{P}_k$ sequentially. The next prediction of the measurement can be obtained as $\mathbf{H}_{k+1}\mathbf{m}_{k+1}^-$ if there is a misdetection of $\mathbf{y}_{k+1}$, $k+1 \notin \mathcal{K}$. However, there is no correction step at $k+1$ due to the lack of measurement which can lead to growing covariance matrix $\mathbf{P}_{k+1}$.

(2) Smoother-based Estimation

RTS smoother is the smoothing method of estimating the current state given the whole measurements instead of just using the current measurement and the previous state. Because on-time measurement is not required in a shake table test, a short-time delay (few seconds or minutes) is allowed and can be used for post processing. Therefore, the study considered using the available measurement not that were not just prior to the current steps ($1 \le i \le k$ & $i \in \mathcal{K}$) but also after the current steps ($k+1 \le i \le T$ & $i \in \mathcal{K}$). RTS smoother, which is similar to but not same as the backward algorithm of KF, is used to estimate the missed measurement $\mathbf{y}_k =?$ ($k \notin \mathcal{K}$) using more measurement in future ($\mathbf{y}_i, k+1 \le i \le T$ & $i \in \mathcal{K}$) in addition to KF. In the study, the KF and RTS smoother results would be combined with expectation-maximization to estimate the dynamic state and missing measurement in the shake table tests. Unlike normal RTS smoother that uses all the measurements for all time steps, this study would impose a constraint on measurement because only partial measurements ($\mathbf{y}_i, 1 \leq i \leq T \; \& \; i \in \mathcal{K}$) could be provided for smoothing.

The close form smoothing solution for a RTS smoother is:

$$p(\mathbf{x}_k|\mathbf{y}_{1:T}) = \mathbf{N}(\mathbf{x}_k|\mathbf{m}_k^s, \mathbf{P}_k^s) \tag{16}$$

where $\mathbf{m}_k^s, \mathbf{P}_k^s$ are the estimated mean and covariance of the current state $\mathbf{x}_k$ based on the whole measurements $\mathbf{y}_{1:T}$.

RTS smoother allows one to refine estimates of current states using the information provided by later observations. The equations for the backward recursion for RTS smoother include the prediction step:

$$\mathbf{m}_{k+1}^- = \mathbf{A}_k \mathbf{m}_k \tag{17}$$

$$\mathbf{P}_{k+1}^- = \mathbf{A}_k \mathbf{P}_k \mathbf{A}_k^{\mathrm{T}} + \mathbf{Q}_k \tag{18}$$

where $\mathbf{m}_k \in \mathbb{R}^{n\times 1}$ and $\mathbf{P}_k \in \mathbb{R}^{m\times m}$ are the mean value and the covariance matrix of the state $\mathbf{x}_k$ computed by the KF.

The correction step in the recursive computation includes:

$$\mathbf{m}_k^s = \mathbf{m}_k + \mathbf{G}_k(\mathbf{m}_{k+1}^s - \mathbf{m}_{k+1}^-) \tag{19}$$

$$\mathbf{P}_k^s = \mathbf{P}_k + \mathbf{G}_k(\mathbf{P}_{k+1}^s - \mathbf{P}_{k+1}^-)\mathbf{G}_k \tag{20}$$

where $\mathbf{m}_k^s \in \mathbb{R}^{n\times 1}$ and $\mathbf{P}_k^s \in \mathbb{R}^{m\times m}$ are the mean and the covariance matrix of the measurement $\mathbf{x}_k$ computed by the RTS smoother. $\mathbf{G}_k = \mathbf{P}_k \mathbf{A}_k^{\mathrm{T}} (\mathbf{P}_{k+1}^-)^{-1}$ is the smoother gain for the correction.

4. Experiment: Shake Table Test

To evaluate the proposed multi-vison approach and algorithm for displacement monitoring, a shake table test was carried out on a three-story aluminum frame (**Figure 5**). Chirp excitation was used as the input ground excitation to induce different levels of structural responses. The three-story aluminum frame (**Figure 5**a) were fabricated with the same story heights of 230 mm. The width and length of the floors in X and Y directions are 202 mm and 204 mm, respectively. The detailed views of the column to floor connection are shown in **Figure 5**b with X direction as the weak direction and Y direction as the strong direction. The center-to-center distances between the two adjacent columns are 149.30 mm in X direction and 178.2 mm in Y direction, respectively. A steel plate with the same mass, 0.66 kg, was affixed to the center of each floor. A shake table (Quanser Shake Table II) was utilized to provide lateral excitation with a payload area size of 460 mm × 460 mm. The maximum stroke limit of the actuator is ±76.2 mm and the frequency range of input motion is 0.5-10 Hz. The proposed approach and LVDT were used to measure dynamic displacements. Numerical simulation, experimental setup, and feature detection (without GLDD implementation) are presented as follows.

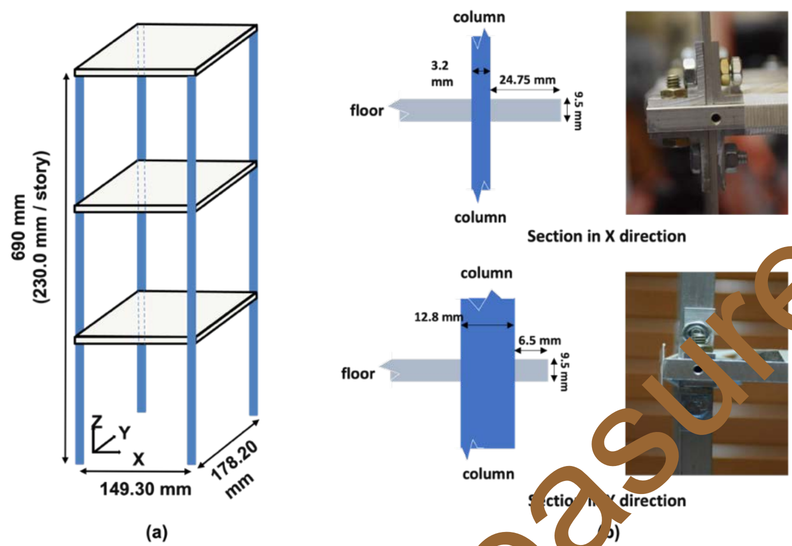


Figure 5. (a) 3D schematic view of the aluminum frame and (b) detailed views of the column-floor connection in X and Y directions.

4.1. Finite Element Simulation

A finite element (FE) model of the same aluminum frame was developed and analyzed using OpenSees (Mazzoni et al., 2006) to understand the structural behavior of the physical model. The same geometry (**Figure** *5*) was used to design the FE model and the columns and floors are modeled by assigning fiber sections to dispBeamColumn and ShellMITC4 elements in OpenSees, respectively. The mechanical properties of the aluminum material for the simulation were: yield strength = 2.5e8 N/m^2, modulus of elasticity = 6.9e10 N/m^2, Poisson's ratio = 0.33, density = 2700 kg/m^3. Following the experimental model, lumped masses of 0.66 kg were assigned to all the three stories. The first three modal frequencies in the X direction of the FE model were 5.68 Hz, 16.06 Hz, and 23.48 Hz (see **Table** ***1***) based on the modal analysis. Ground excitations using an upchirp excitation (designed as 0.5-4.5 Hz) and the free vibration after the excitation are simulated on the FE model. By knowing the modal parameters (e.g., modal frequencies) from FE analysis, the ground excitation can be well designed for the real experiment to cover different frequency spectrums while maintaining the safety during the laboratory test.

Table 1. Natural frequency (first three modes in X direction) and MAC value comparison between FEM and output-only system identification results using (virtual) free vibrations and experimental measurements.

Meas.	Sys. ID Method	Mode1 Freq. (Hz)	Mode1 MAC	Mode2 Freq. (Hz)	Mode2 MAC	Mode3 Freq. (Hz)	Mode3 MAC
Virtual (FEA)	Modal Analysis	5.68	-	16.06	-	23.48	-
	FDD	5.62	1.00	15.87	1.00	23.19	1.00
	SOBI	5.62	1.00	15.87	0.96	23.07	[illegible]
	SSI	5.60	1.00	15.85	1.00	23.09	1.00
Experiment (CV)	FDD	5.15	-	14.95	-	[illegible]	-
	SOBI	5.15	-	14.99	-	24.59	-
	SSI	5.14	-	15.02	-	24.55	-

4.2. Experimental Setup

Twenty-three AprilTags of "25h9 tag family" with unique IDs were attached to the aluminum frame. Eight tags were attached to the key end locations on the front surface of frame's floors to record displacement time histories of the structure during the experiment. The remaining seventeen tags were attached onto the surface of the table base to perform camera location/pose estimation and to record the displacement time history of the base. Moreover, the displacement and acceleration time histories of the base were recorded by the LVDT and accelerometer integrated with the shake table.

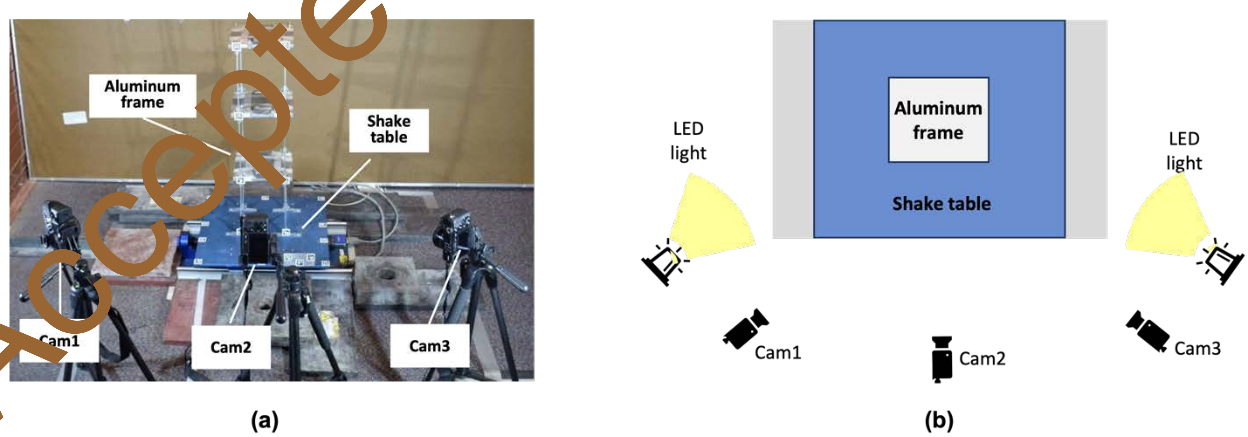


Figure 6. (a) Photo and (b) schematics of the experimental setup of the shake table test with the aluminum frame and cameras.

The aluminum frame (**Figure 6**a) was subjected to the same upchirp (0.5-4.5 Hz) excitation (**Figure 7**) as in the FE simulation (**Figure 5**) with the maximum base displacement of ±10 mm and the duration of 15 seconds. Three portable cameras (**Figure 6**) were used to monitor the shake table tests: two Sony α6400 cameras (denoted as Cam1 and Cam2) and one Sony α6000 camera (denoted as Cam3). Multiple DC light-emitting diode (LED) sources were deployed further away from the shake table for balanced illumination condition. Please note that there is a trade-off between different (camera) angles of view: wide-angle allows more context in view while sacrificing the density of pixels over foreground (structures); narrow angle allows denser pixels over structures while covers less area. Based on the size of the three-story frame structure, a medium field view (f=16 mm) was adopted with field of view angles of 72.59° in vertical and 52.27° in horizontal directions. Camera parameters were set as the same for the three cameras to record high-quality (1920 x 1080 pixel^2) videos of the structure in vibration: focal length = 16 mm, ISO = 2000 (the sensitivity to light), frame rate = 59.94 fps, shutter speed = 1/165 s (6.1 ms), and camera aperture = f6.3.

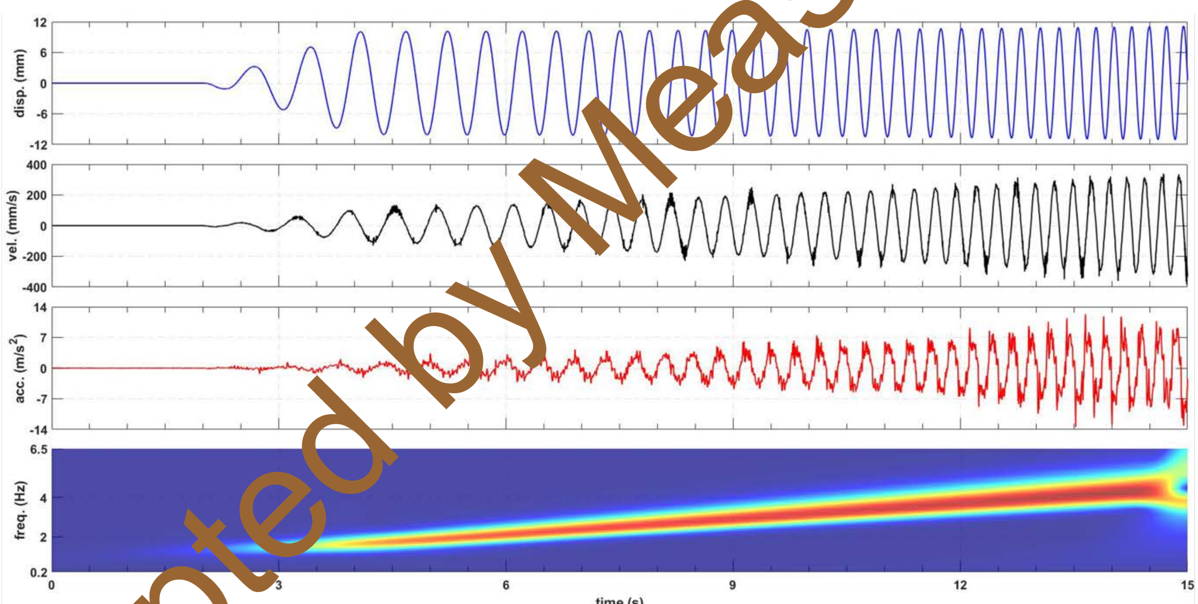


Figure 7. Time histories of (a) the base displacement of the chirp ground excitation (measured by LVDT), (b) velocity using 1st order differentiation, (c) acceleration using 2nd order differentiation, and (d) the wavelet transform of the base displacement.

Ambient soundtracks of the test were used for video matching between the three cameras using cross correlation method to compute the differences in time (**Figure 8**). Cam1-audio was used as a reference and the time shifts were +2.54 s (+152 frames) and +2.59 s (+155 frames) for the Cam2-audio (-video) and Cam3-audio (-video) channels, respectively. After the synchronization of frames from multiple perspectives, muti-vision triangulation is performed.

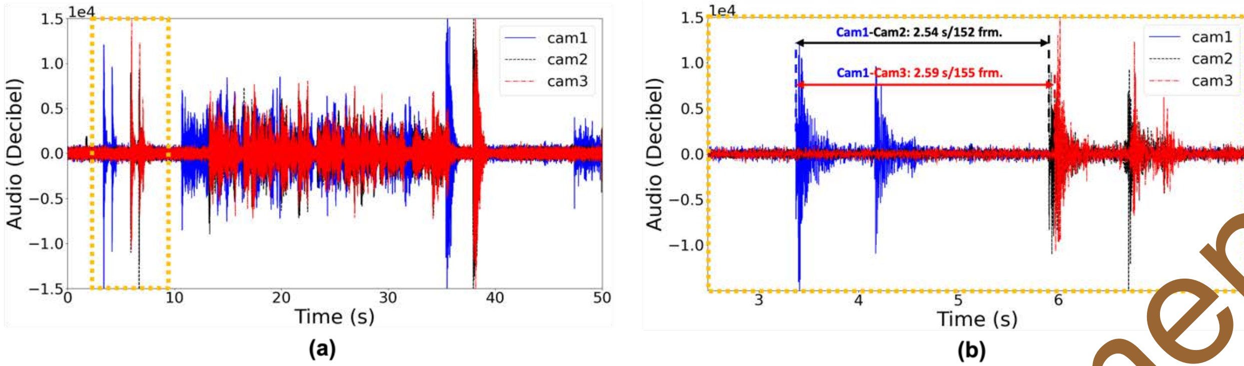


Figure 8. (a) Synchronization processing using the audio data from the three cameras and cross-correlation, and (b) the detailed view within the time window of (2.5-8 s).

4.3. Tag Detection without GLDD

Evaluation of camera calibrations of the three cameras (**Figure 9**) were conducted with the blue dots representing the 3D location in WCS projected onto sample images (using recognized camera parameters) and red circles representing the detected location in SCS. The average error in 2D SCS between the detection locations $[\hat{u}_i, \hat{v}_i]^T$ using AprilTags and the projected locations in 2D SCS were 2.99 pixel for Cam1, 3.53 pixel for Cam2, and 2.97 pixel for Cam3, respectively. The average errors in 3D SCS between the computed location $[\hat{X}_i, \hat{Y}_i, \hat{Z}_i]^T$ and the measurements by rulers in 3D WCS were 2.03 mm for Cam1, 1.58 mm for Cam2, and 1.72 mm for Cam3, respectively.

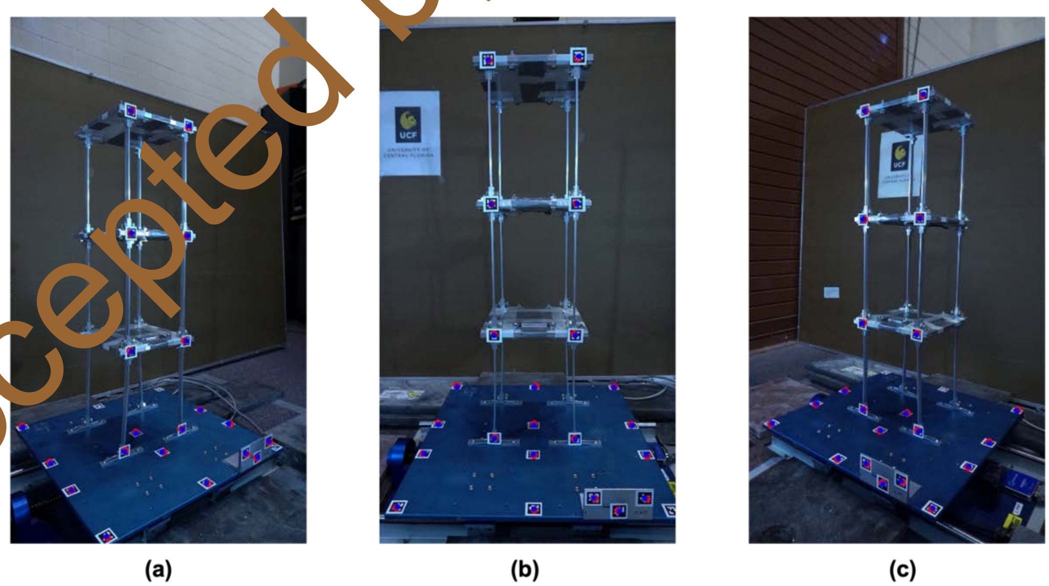


Figure 9. Evaluation of camera calibration and pose estimation in the small-scale shake test. (note: red circles are projection using camera parameters and the ground truth locations in WCS and blue dots are detected points in SCS)

The detection performance is shown in **Figure *10*** for the three cameras. A successful detection event by raw tag detection technique is denoted as a gray solid circle and a failed detection event is denoted as a red circle for each of the three perspectives at each synchronized frame/time across the three cameras. The detection performances are shown for different floors from top to bottom: T0/T1 on the 3^{rd} floor, T2/T3 on the 2^{nd} floor, T4/T5 on the 1^{st} floor, and T6/T7 on the base. When the motion blur was little (0-12 s, 0-720 frames), the detection performance for the raw tag detection is satisfactory with all the tags on the frame successfully detected and the success rate was 100%. As the vibration becomes large enough to cause mild motion blur (12.3-13.2 s), the tags on the 3^{rd} floor (T0 and T1) were difficult to identify. As the excitation frequency increased (13.2-14.2 s), tags on the 3^{rd} floor more frequently failed to be detected and tags on the 2^{nd} floor experience misdetections. During the last one second, when the excitation frequency was close to the 1^{st} natural frequency of the structure, even tags on the 1^{st} floor were difficult to be detected just using the raw tag detection technique. T6 and T7 on the base floor showed a 100% detection rate for all the cameras.

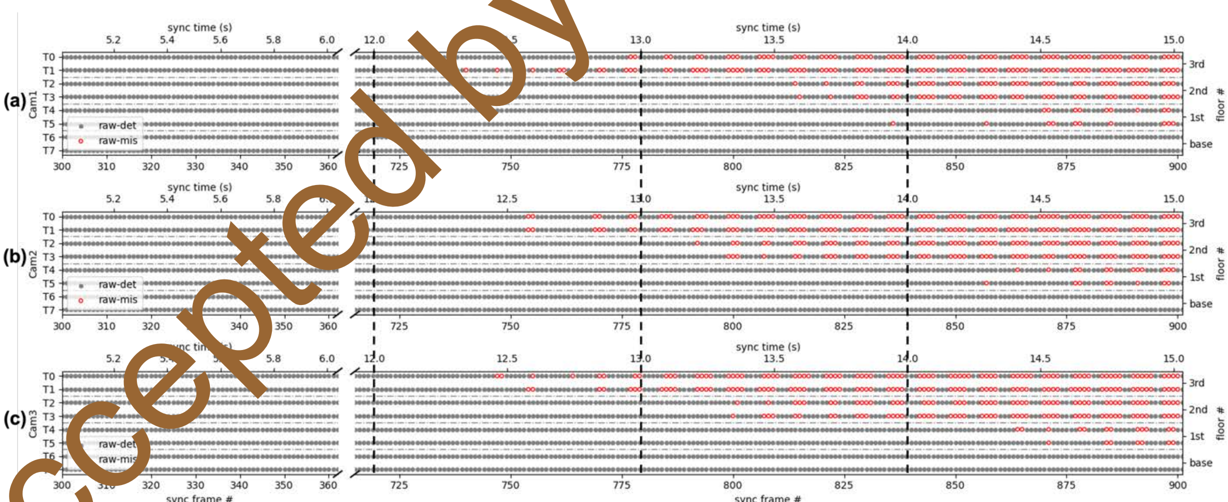


Figure 10. Tag detection evaluation based on raw image frames from (a) Camera 1, (b) Camera 2, and (c) Camera 3 in the shake table test.

The vertical lines in **Figure *10*** delineate the time segments which are associated with different degrees of motion blur. As shown in **Table *2***, the missed rates for the mild level of motion blur (12-13 s) were 12/480, 13/480, and 13/480 for Cam1-3, respectively. The missed rates for the medium level of motion blur (13-14 s) increased to 70/480, 91/480, and 94/480 for Cam1-3, respectively. The missed rates for the severe level of motion blur (14-15 s) were 151/480 (Cam1), 162/480 (Cam2), and 159/480 (Cam3). Misdetection events occur when the velocities of the tags were higher on upper floors. In order to obtain more dynamic measurements, GLDD is needed to enhance the detection rate.

5. Result and Discussion

5.1. Augmented Detection with Multi-Vision and GLDD

As an example, the image-based GLDD process on a video frame is shown in **Figure *11***. When there was motion blur on images, the detection algorithm using the raw blurred image (**Figure *11***a) could not identify all the tags on the structure, in contrast to a static frame without motion blur (**Figure 9**b). However, the detection rate can be improved by using the GLDD module to an extent. For example, the GDD algorithm could improve the detection on the front surface of the frame (**Figure *11***b), especially on the 2nd floor (**Figure *11***f) that was at the center of the whole image. However, while the GAN-based deblurring on the whole image could only improve detection in a certain focused area (e.g., center), the global image restorage was not enough to make the tags on the 3rd floor detectable (**Figure *11***h). In addition, the objects that were off centered (e.g., base) did not necessarily become sharper using the GAN-based deep deblurring method, which could even impose adverse effects making previously detected tags less detectable (**Figure *11***d) after the image restoration. However, the local deblurring on a local crop image (**Figure *11***j) could shift more attention on the important features (e.g., tags, structural features) making the bit features sharper for successful detection (**Figure *11***j). One may argue that this improvement might be due to the cropped image size. Experiment result (**Figure *11***i), though, proved that using an image crop even around a tag would not necessarily improve success rate.

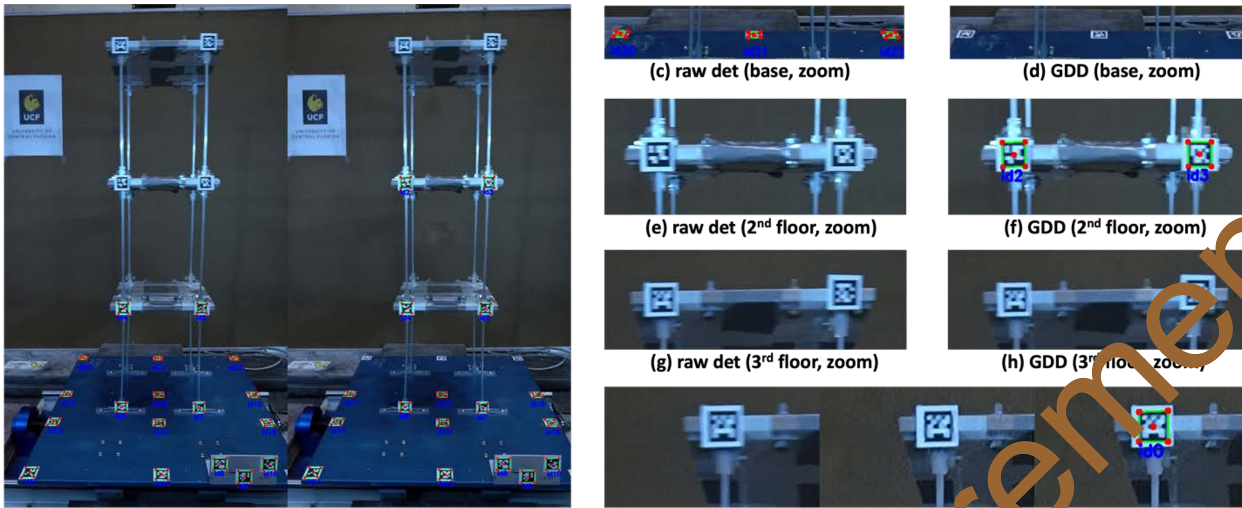


Figure 11. GLDD process steps on one video frame: (a) detection result on raw image, (b) detection result on globally deblurred image, (c) raw detection result and (d) GDD processed result on the table base, (e) raw detection results and (d) GDD processed result on the 2nd floor, (g) raw detection result and (h) GDD processed result on the 3rd floor, (i) detection result on the local image near Tag0, (j) its locally deblurred image, and (k) LDD processed result.

The limit of the proposed framework was evaluated using the upchirp excitation whose actual frequency window is about 0.5-6.2 Hz covering the 1st natural frequency of the structure (5.15 Hz in **Table** *1*) near the end of the experiment. Because the transient displacement input on the table base included frequency components at the 1st natural frequency near 15 s (**Figure** *7*), the short-time resonance caused the frame to shake violently and resulted in severe motion blurs in videos. As shown in **Figure** *12*, the detection performance for the GLDD process for each of the camera is presented with blue-plus symbols denoting successful detections using the GDD and red-cross symbols denoting successful detection using the local detection and deblurring (LDD). The detailed performances of GDD and LDD were compared using an ablation study of the different augmentation strategies (**Table** *2*). When the time was 12-13 s in the shake table test, there were 12 misses, 13 misses, and 13 misses for Cam1-3, respectively. With the GDD, the miss counts went down to 4, 2, and 10 for the three cameras, respectively. With the additional LDD, the miss counts went further down to 2, 0, and 0 for the three cameras, respectively, making the total restoration rates of 10/12 (83.3%) for Cam1, 13/13 (100.0%) for Cam2, and 13/13 (100.0%) for Cam3. The average restoration rate for the three cameras were 35/38 (92.1%) for mild-level motion blur. When the excitation frequency increases from around 3.9 Hz to 4.2 Hz during 13-14 s (**Figure** *7*d), the restoration rates for the GLDD process were 36/70 (51.4%) for Cam1, 54/91 (59.3%) for

Cam2, and 39/94 (41.5%) for Cam3. The average restoration rate for the three cameras was 129/255 (50.6%) for the medium-level motion blur. During the last one second (14-15 s) when frequency span of the excitation (3.9-6.2 Hz) overlapped with the 1st natural frequency of the structure (5.15 Hz), the restoration rates by GLDD decreased to 44/151 (29.1%) for Cam1, 49/162 (30.2%) for Cam2, and 26/159 (16.4%) for Cam3. The average restoration rate for the three cameras was 119/472 (25.2%) for severe-level motion blur.

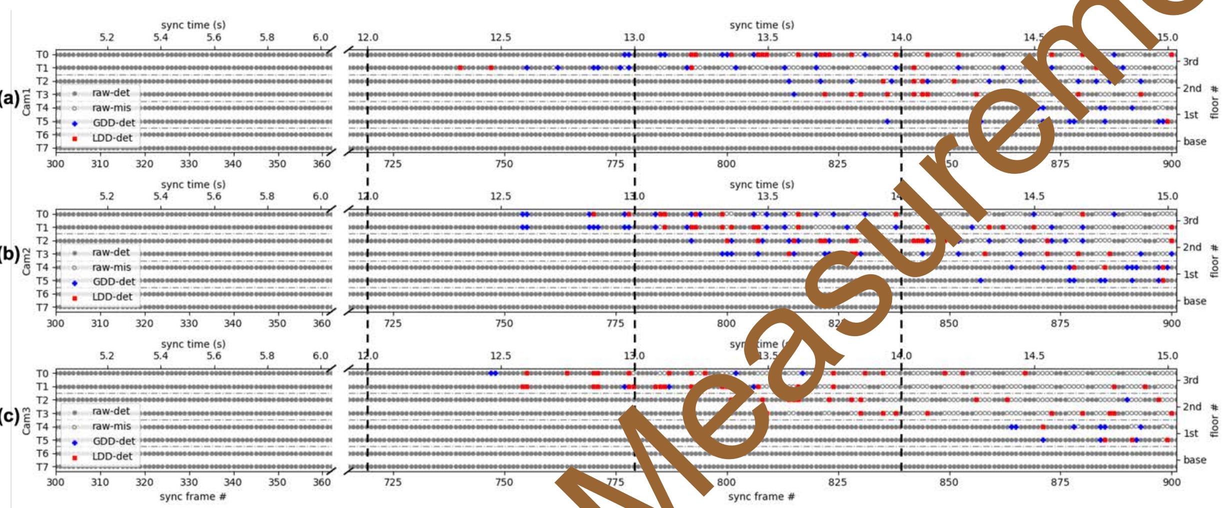


Figure 12. Tag detection evaluation with GLDD processing from (a) Camera 1, (b) Camera 2, and (c) Camera 3 in the shake table test.

After analyzing all the video frames during the whole shake table test (0-15.3 s, 0.5-4.5 Hz), it was found that the restoration rates of the GLDD were 92/243 (37.9%) for Cam1, 126/281 (44.8%) for Cam2, 82/284 (28.9%) for Cam3. The different performances among cameras were due to the relative location and pose of cameras with respect to the shaking aluminum frame showing the effect of the camera placement on achieving high quality CV-based results. The GLDD process itself restored 94/207 (45.4%) of previous misdetections using raw images. The multi-vision strategy did take effect in the detection augmentation as well. Take Cam3 for example, the missed counts were brought from 13 down to 6 during 12-13 s, from 94 down to 60 during 13-14 s, and from 159 to 133 during 14-15 s. For the whole test (0-15.3 s), the total miss count for Cam3 was brought from 284 down to 207 with a 26.0% drop. With the implementation of both strategies (multi-vision and GLDD), the total miss count for the experiment is brought down to only 113 (**Table *2***) with 75.0% measurements retrieved (from the previous misdetections) compared to just using raw images from one single camera (e.g., Cam3).

Table 2. Hit and miss counts using different tag detection methods for each camera and hybrid setting during different time windows of the shake table test.

detection method		t: 12-13s (frm: 719-778)				t: 13-14s (frm: 779-838)				t: 14-15s (frm: 739-899)				t: 0-15.3s (frm: 0-[illegible])			
		cam1	cam2	cam3	multi	cam1	cam2	cam3	multi	cam1	cam2	cam3	multi	cam1	cam2	cam3	multi
raw	det	468	467	467	474	410	389	386	420	329	318	321	347	7133	709[illegible]	709[illegible]	7169
	mis	12	13	13	6	70	91	94	60	151	162	159	133	243	281	284	207
GDD	det	476	478	470	480	428	418	395	436	357	349	331	367	7[illegible]87	7171	7115	7212
	mis	4	2	10	0	52	62	85	44	123	131	149	11[illegible]	189	205	261	164
GLDD	det	478	480	480	480	446	443	425	461	373	367	347	390	[illegible]225	7221	7174	7263
	mis	2	0	0	**0**	34	37	55	**19**	107	113	133	**90**	151	155	202	**113**

5.2. CV-based Displacement Measurement

The pixel coordinates of the tag centers were localized on the collected video frames. For example, the image-based detection results (from Cam1) for the four floor levels (T0-3rd floor, T2-2nd floor, T4-1st floor, T6-base) are shown in **Figure** *13*. Results from raw-image detection are denoted as gray dots, additional results from GDD process are denoted as blue "+" symbol, and the additional results from LDD process are denoted as red "×" symbol. From bottom to top of the structure, the increasing motion blur made it more and more difficult to detect using raw images. The GDD retrieved almost all the misdetections (denoted as "miss" in the study) on the 1st floor (**Figure** *13*c and g) from 14.5-15.0 s (870-900 frames). On the 2nd floor and 3rd floor, the GLDD performed well from 12.8-13.8 s (775-825 frames) by restoring all the misdetections from raw images. LDD were more robust dealing with challenge events from 13.8-15.1 s (825–900 frames) when the motion blur was at a severe level.

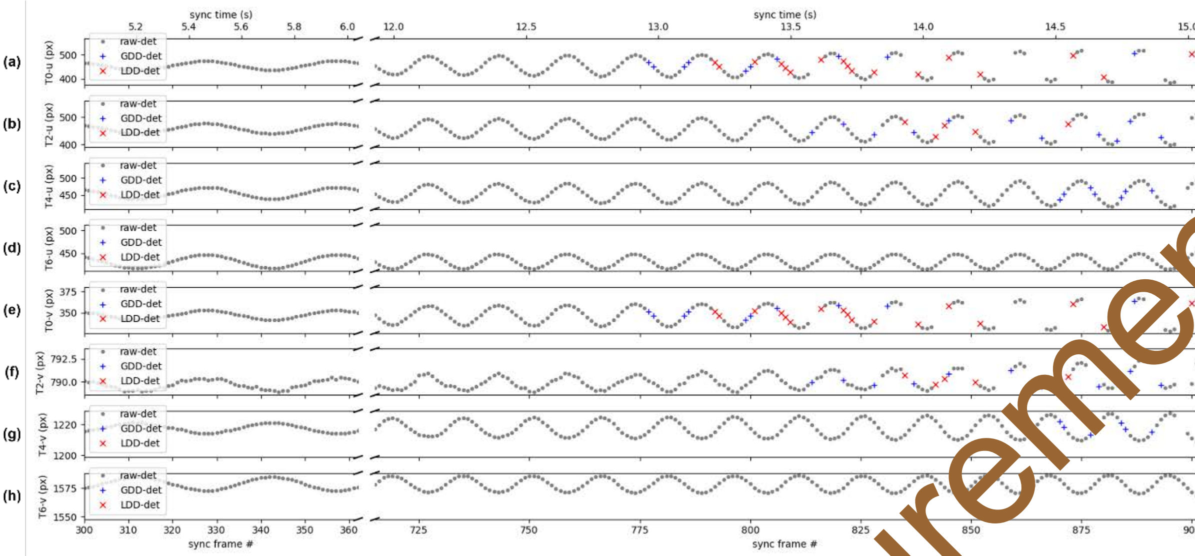


Figure 13. Localization results of (a-c) u and (e-h) v sensor coordinate system for Tag0 on 3rd floor, Tag2 on 2nd floor, Tag4 on 1st floor, and Tag6 on base using GLDD module.

The time histories of displacement were compared (**Figure** *14*) among the designed displacement input, LVDT measurement, single-vision measurements, and dual-vision measurements on the table base. There was a small difference (0.506 mm) between the designed (displacement) input and the LVDT measurement in the shake table test. This study used LVDT measurement as the baseline for comparison among vision-based measurements using root-mean-square error (RMSE). RSMEs between single-vision methods and LVDT were 0.881 mm (Cam1), 0.829 mm (Cam2), and 0.222 mm (Cam3), respectively. RSMEs between dual-vision methods and LVDT were 0.823 mm (Cam1-2), 0.507 mm (Cam1-3), and 0.949 mm (Cam2-3), respectively. The measurements from single-vision and dual-vision matched well with the LVDT measurement. In addition, the measured displacements were found consistent for single–vision setting (**Figure** *15*) and dual-vision setting (**Figure** *16*) on different floors with different extent of motion blur, validating the measurement robustness using multiple perspectives. Dual-vision improved the confidence in measurement compared with single-vision, although the accuracies of the two were similar in this study. When using single-vision, a strong assumption is required that allows only in-plane movement. The reason that both dual-vision and single-vision achieved similar accuracy in our study is that the excitation only caused in-plane vibration, meeting the required assumption for single-vision measurement. The implementation of augmentation strategies (GLDD and multi-vision) could address mild- and medium-level motion blur. However, when the motion blur is too severe (e.g., vibration near natural frequencies), data fitting is needed to supplement and help interpolate/estimate the missing measurement.

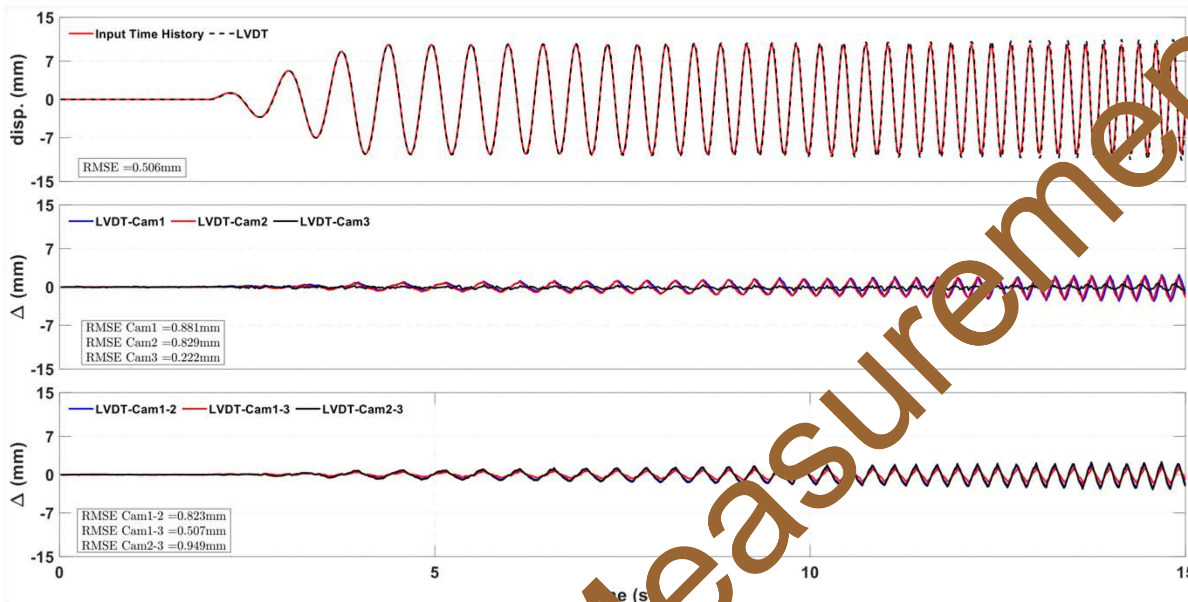


Figure 14. (a) Time histories of designed displacement input and LVDT measurement, (b) the differences between single-vision measurements and LVDT, and (c) the differences between dual-vision measurements and LVDT.

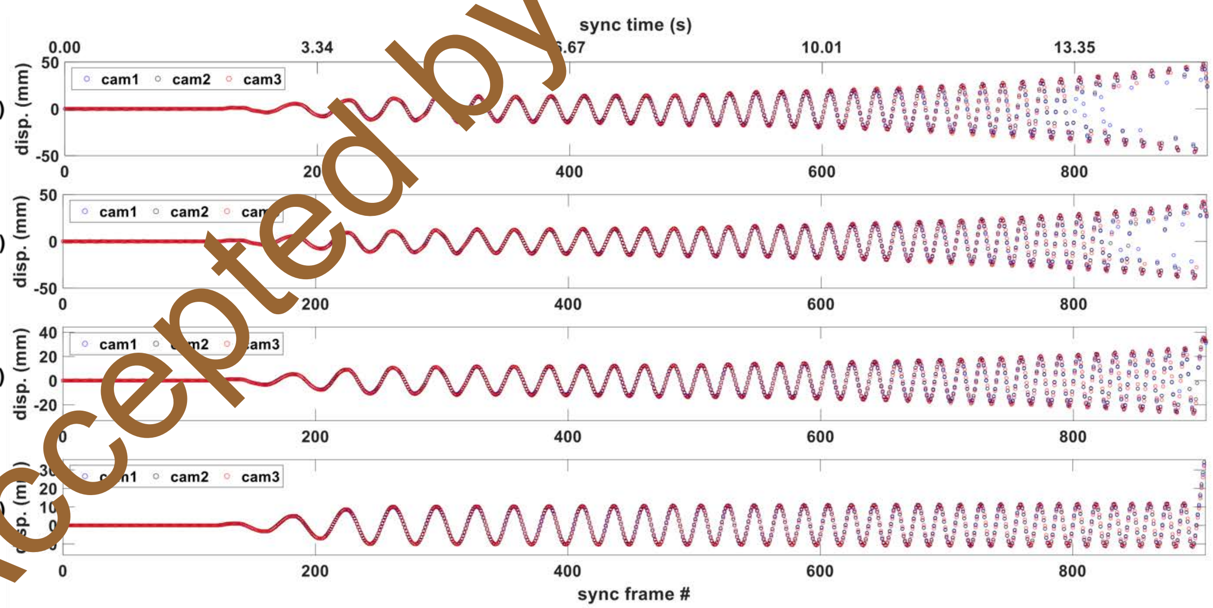


Figure 15. Displacement measured by single-vision method on (a) 3rd floor, (b) 2nd floor, (c) 1st floor, and (d) base.

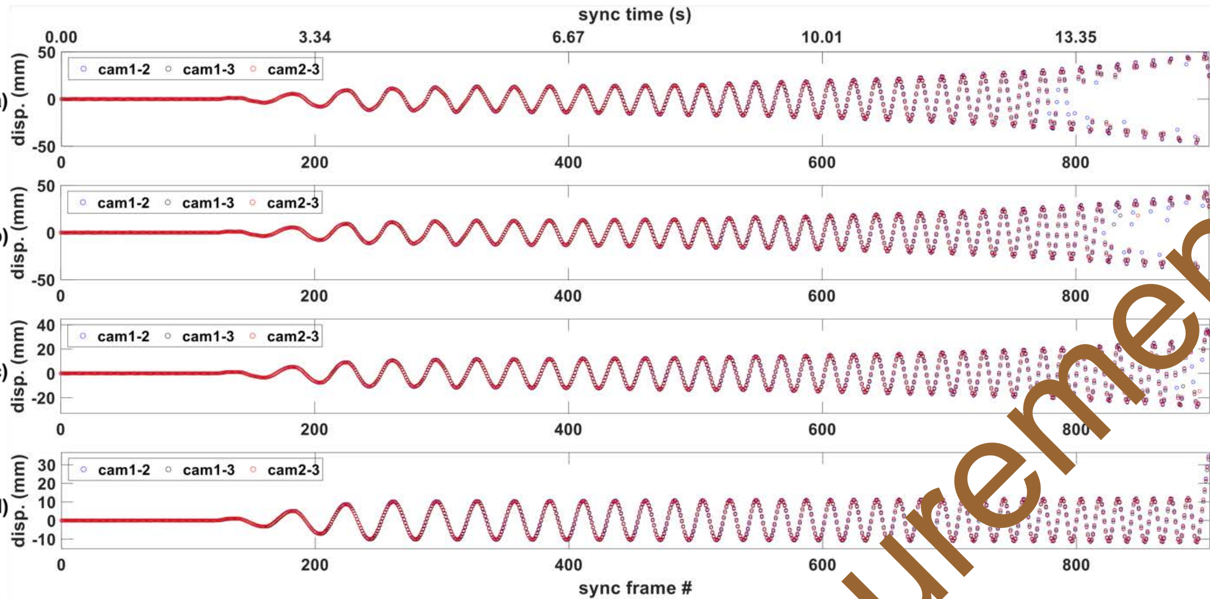


Figure 16. Displacement measured by dual-vision method on (a) 3rd floor, (b) 2nd floor, (c) 1st floor, and (d) base.

5.3. Data Fitting with Filter and Smoother

When motion blur was at mild or medium level in the experiment, the proposed GLDD method could restore some images for feature detections but could not resolve severe image blurs. Excessive motion blur was studied using filtering and smoothing methods to estimate the missed measurements. The virtual measurement on the 3rd floor of the FE model was used to evaluate the measurement estimation. Two virtual incomplete measurements between the two thresholds ($|d_1| \leq$10 mm as shown in **Figure 17**a-b, and $|d_2| \leq$15 mm as shown in **Figure 17**c-b) were masked with (k, i) within the time window of 12 s<t_k<15 s. The masked observations were treated as failed/missed observations $((?_k)_{\mathrm{i}}, (k, i) \in \mathcal{M})$. If the number of measurements for a single degree of freedom system is one, *i* can be dropped and the failed/missed observation can be presented as $(?_k, k \in \mathcal{M})$. Incomplete measurements were used to restore the unknown observation. The sampling time was dt = 0.0167 s to simulate vision-based measurement. The missed (virtual) measurements were within 12-15 s excluding the raw detections and the GLDD detection, to simulate the actual misdetection caused by the severe motion blur. In the KF and smoother setting, the initial system matrix was set as $\mathbf{A}_0 = [1, dt; 0,1]$, the initial guess of the state was set as $\mathbf{x}_0 = [0,0]^T$, and control was not considered in process equation. The covariance matrix of the process $\mathbf{Q}_0$ was set as [0.5, 0; 0, 0.5], and the dynamic model/transition matrix was set as $\mathbf{R}_k = 0.1$ mm^2 based on the RMSE's in the evaluation of vision-based methods.

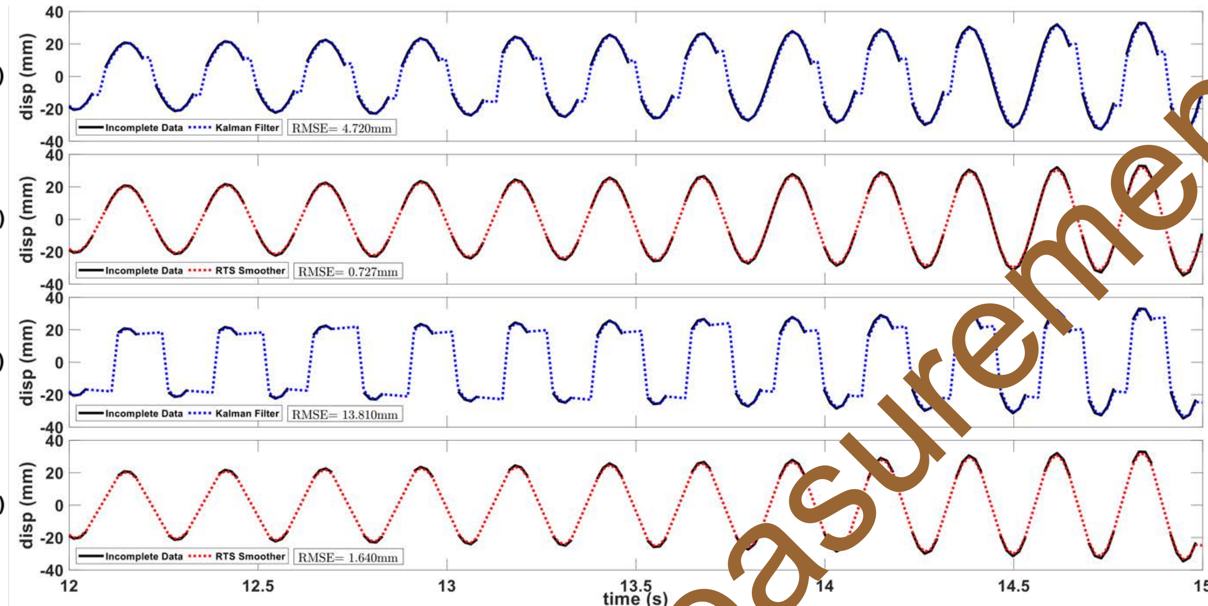


Figure 17. Data fitting performance of Kalman filter and RTS smoother in two scenarios with (a-b) medium-level and (c-d) severe-level of (virtual) measurements from FE analysis.

Figure *17* shows the performance of measurement fitting using KF and RTS smoother within the two incomplete time history data masked from 12-15 s. When there was a small number of missed measurements, e.g., 12.22% misses among the local time window of 12-15 s as 2.44% misses among the whole 0-15 s, KF still worked by neglecting the correction step (**Figure *17***a). However, when there were considerable number of missed measurements, e.g., 38.89% misses within the local time window of 12-15 s as 7.78% misses within the whole 0-15 s, the covariance matrix $\mathbf{P}_k^-$ for state was enlarged without the necessary correction step. It was observed that the larger 13.810 mm RMSE occurred for KF estimation in severer blur case (**Figure *17***c) compared to 4.720 mm RMSE in mild blur case (**Figure *17***a).

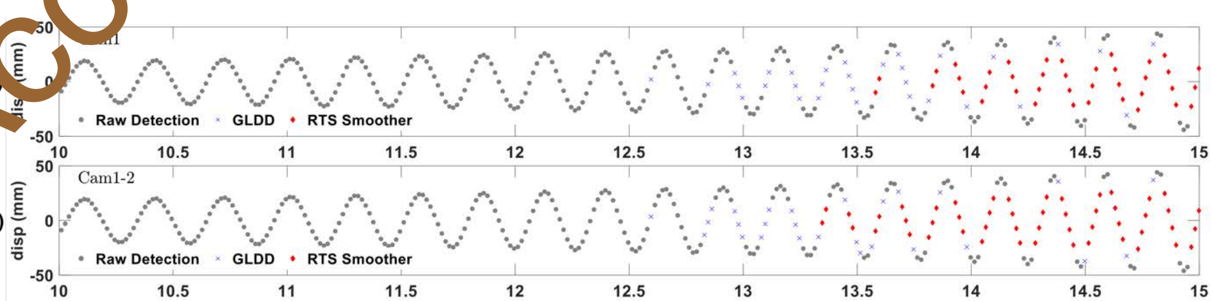
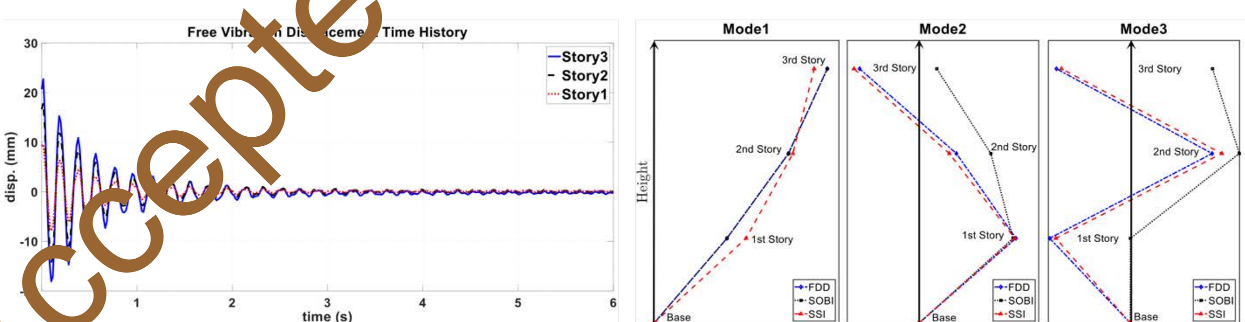


Figure 18. Data fitting performance of RTS smoother in experiment measurements from (a) single-vision (Cam1) and (b) dual-vision (Cam1-2).

The RTS data fitting took all available measurements into consideration, yielding better estimation performance (**Figure** ***17***b-d) with an improved RMSE of 0.727 mm and 1.640 mm for both mild and sever cases. The proposed RTS smoother-based fitting method was also implemented on the CV-based displacement measurements on the 3rd floor in the actual shake table test. The parameters of the RTS smoother for the experimental measurements were chosen as the same as the virtual one. As shown in **Figure** ***18***, the measurement from the raw detection and the measurement from the GLDD are denoted as gray dots and blue crosses, respectively. The estimation using the RTS smoother in single-vision case (**Figure** ***18***a) and dual-vision case (**Figure** ***18***b) are denoted as red diamond symbols. The data fitting results showed satisfactory estimation using RTS smoother.

5.4. Application of System Identification

After augmentation from the deblurring module and RTS smoother-based data fitting module, the measurement result can be used for system identification providing modal information for future applications (e.g., modal updating, structural damage identification). System identification can be based on structural displacements, such as free, forced, or ambient vibrations. To demonstrate the application of CV-based measurements in system identification, three output-only system identification methods of frequency domain decomposition (FDD) (Brincker et al., 2001), second-order blind identification (SOBI) (Belouchrani et al., 1997), and stochastic subspace identification (SSI) (Van Overschee & De Moor, 2012) were compared using both the virtual (from FE analysis) and experimental free vibration displacements.



(a) Cam1-2 (b) Mode1 (c) Mode2 (d) Mode3

Figure 19. (a) Free-vibration displacements measured by dual-vision CV, and (b-d) the first three modal shapes from the system identification using FDD, SOBI, and SSI methods.

Table *1* shows that the identified modal frequencies of the first three modes match well with the FEM modal analysis from OpenSees. Modal assurance criterion (MAC) values for all three modes are above 0.96 (except SOBI for mode-3 with 0.53) indicating a satisfactory performance of FDD and SSI compared to SOBI. System identifications were also performed using the experimental displacements (e.g., free vibrations in **Figure *19***a) measured by proposed method. The identified modal frequencies (see **Table *1***) are very close with the 1st frequency identified as 5.15 Hz (FDD), 5.15 Hz (SOBI), and 5.14 Hz (SSI). The average differences between the identified modal frequencies using FDD, SOBI, and SSI with the FE-based modal frequencies are 9.4%, 6.7%, and 4.3% for modes 1, 2, and 3, respectively. **Figure *19***b-d show the first three mode shapes using the experimental measurements. It is found that the mode shapes between FDD and SSI are close to each other, especially for mode-2 and mode-3. In general, the identification results were consistent among the three methods using proposed multi-vision method.

6. Conclusion

The study proposed a multi-vision monitoring approach using low-cost cameras to measure structural displacements in shake table tests with the augmentations from novel application of deep learning-based image deblurring and Rauch-Tung-Striebel (RTS) Smoother. The proposed global-local deblurring and detection (GLDD) module was able to restore clearer images for feature detection, especially when dealing with mild-level motion blur with average restoration rates of 92.1%. The restoration rates dropped to 50.6% for mild-level motion and further to 25.2% for severe-level motion with the increasing severity of image blurs. Misdetections due to excessive motion blur were estimated with filtering and smoothing-based methods using incomplete measurements. RTS smoother is able to achieve a satisfactory data estimation (with a RMSE of 1.64 mm) outperforming Kalman filter (with a RMSE of 13.81 mm) in the scenario with severely incomplete observations. RTS smoother helped accurately estimate missed measurement due to severe blurs, especially when the misdetections were consecutive as typical in shake table tests. Implementation of GLDD module was tested in a shake table test of a three-story aluminum frame and was validated with linear variable differential transformer measurement. Results show the potential of the proposed approach in measuring dynamic displacement. The proposed multi-vision and GLDD strategies can retrieve 75.0% measurements from previous misdetections (by just using raw images from one single camera) and the data fitting module can complete the rest.

The main contribution of the study includes: (1) proposing a multi-vision displacement measurement approach using low-cost cameras with novel deblurring module and RTS smoother-based data fitting module to address the motion blur issue; (2) studying the effectiveness of the modules in dealing with different levels of motion blur in shake table tests; (3) providing the guidelines for using the proposed approach in shake table tests and the augmented displacement that can be used in the further structural analysis. The proposed method can be employed in a range of other applications (e.g., structural dynamics, finite element model updating) and be extended to real-world applications, such as deflection measurement of bridge due to traffic loads, vibration monitoring on high-rise buildings in earthquakes, and monitoring of relative displacement between key structural members (e.g., inter-story, beam-column joints). One limitation of the study is that: although the proposed multi-vision scheme and deblurring module is found to retrieve misdetections due to mild and median motion blur, but it cannot restore images from excessive motion blur, which is still a challenging issue for image processing. In addition, the effectiveness of the proposed method using natural structural features and under challenging environmental conditions (e.g., poor illumination, occlusion of features) remains to be studied due to the limited extent of this work. Future works will focus on studies of, such as effects of challenge conditions (e.g., illumination, occlusion) in real application scenarios, faster algorithm on displacement estimation, and error analysis of sensor deployment of the multi-vision system.

Acknowledgment

The authors thank Dr. Kevin Mackie from the University of Central Florida for providing the shake table facilities. Help from Mr. Mostafa Iraniparast, Mr. Seyed Sina Shid-Moosavi, Ms. Lisa Barra, and Ms. Aya Leone in the experiments is appreciated. This research is partially supported by the Florida Department of Transportation (grant no.: BDV24 562-14 and BED26 562-4) and the Florida Department of Environmental Protection William W. "Bill" Hinkley Center for Solid and Hazardous Waste Management (award no.: AWD08952, project no.: P0184923).

Author Contribution:

The authors confirm contribution to the paper as follows: study conception and design: P. Sun; data collection: M. Vasef; analysis and interpretation of results: P. Sun, M. Vasef; draft manuscript preparation and revision: P. Sun, M. Vasef, L. Chen. All authors reviewed the results and approved the final version of the manuscript.

References

Abdel-Aziz, Y. I., Karara, H. M., & Hauck, M. (2015). Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. *Photogrammetric engineering & remote sensing*, *81*(2), 103-107.

Belouchrani, A., Abed-Meraim, K., Cardoso, J.-F., & Moulines, E. (1997). A blind source separation technique using second-order statistics. *IEEE Transactions on signal processing*, *45*(2), 434-444.

Bezcioglu, M., Yigit, C. O., Karadeniz, B., Dindar, A. A., El-Mowafy, A., & Avcı, Ö. (2023). Evaluation of real-time variometric approach and real-time precise point positioning in monitoring dynamic displacement based on high-rate (20 Hz) GPS Observations. *GPS Solutions*, *27*(1), 43.

Brincker, R., Zhang, L., & Andersen, P. (2001). Modal identification of output-only systems using frequency domain decomposition. *Smart Materials and Structures*, *10*(3), 441.

Choi, H.-S., Cheung, J.-H., Kim, S.-H., & Ahn, J.-H. (2011). Structural dynamic displacement vision system using digital image processing. *Ndt & E International*, *44*(7), 597-608.

Chu, A. (2005). CHAPTER 17 - Sensors for Mechanical Shock. In J. S. Wilson (Ed.), *Sensor Technology Handbook* (pp. 457-480). Newnes. https://doi.org/https://doi.org/10.1016/B978-075067729-5/50057-4

del Rey Castillo, E., Allen, T., Henry, R., Griffith, M., & Ingham, J. (2019). Digital image correlation (DIC) for measurement of strains and displacements in coarse, low volume-fraction FRP composites used in civil infrastructure. *Composite Structures*, *212*, 43-57.

Dong, C.-Z., Bas, S., & Catbas, F. N. (2020). A portable monitoring approach using cameras and computer vision for bridge load rating in smart cities. *Journal of Civil Structural Health Monitoring*, *10*, 1001-1021.

Dong, C.-Z., & Catbas, F. N. (2021). A review of computer vision–based structural health monitoring at local and global levels. *Structural Health Monitoring*, *20*(2), 692-743.

Gao, X., & Zhang, T. (2021). *Introduction to visual SLAM: from theory to practice*. Springer Nature.

Greenbaum, R. J., Smyth, A. W., & Chatzis, M. N. (2016). Monocular computer vision method for the experimental study of three-dimensional rocking motion. *Journal of Engineering Mechanics*, *142*(1), 04015062.

Guo, J., Jiao, J., Fujita, K., & Takewaki, I. (2019). Damage identification for frame structures using vision-based measurement. *Engineering Structures*, *199*, 109634.

Han, K. a. W., Yunhe and Tian, Qi and Guo, Jianyuan and Xu, Chunjing and Xu, Chang. (2020). GhostNet: More Features From Cheap Operations. *InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2020 (pp. 1580-1589)*.

Kogut, J. P., & Pilecka, E. (2020). Application of the terrestrial laser scanner in the monitoring of earth structures. *Open Geosciences*, *12*(1), 503-517.

Krogius, M., Haggenmiller, A., & Olson, E. (2019). Flexible layouts for fiducial tags. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),

Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D., & Matas, J. (2018). Deblurgan: Blind motion deblurring using conditional adversarial networks. Proceedings of the IEEE conference on computer vision and pattern recognition,

Kupyn, O. a. M., Tetiana and Wu, Junru and Wang, Zhangyang. (2019). DeblurGAN-v2: Deblurring (Orders-of-Magnitude) Faster and Better. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 8878-8887*.

Liu, Y., Haridevan, A., Schofield, H., & Shan, J. (2022). Application of Ghost-DeblurGAN to Fiducial Marker Detection. 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),

Ma, Y., Soatto, S., Košecká, J., & Sastry, S. (2004). *An invitation to 3-d vision: from images to geometric models* (Vol. 26). Springer.

Ma, Z., Choi, J., & Sohn, H. (2022). Real-time structural displacement estimation by fusing asynchronous acceleration and computer vision measurements. *Computer-Aided Civil and Infrastructure Engineering*, *37*(6), 688-703.

Mazzoni, S., McKenna, F., Scott, M. H., & Fenves, G. L. (2006). OpenSees command language manual. *Pacific earthquake engineering research (PEER) center*, *264*(1), 137-158.

Muralikrishnan, B. (2021). Performance evaluation of terrestrial laser scanners—A review. *Measurement Science and Technology*, *32*(7), 072001.

Olson, E. (2011). AprilTag: A robust and flexible visual fiducial system. 2011 IEEE international conference on robotics and automation,

Ramakrishnan, S. a. P., Shubham and Gangopadhyay, Aalok and Raman, Shanmuganathan. (2017). Deep Generative Filter for Motion Deblurring. *InProceedings of the IEEE international conference on computer vision workshops 2017 (pp. 2993-3000)*.

Rychlicki, M., Kasprzyk, Z., & Rosiński, A. (2020). Analysis of accuracy and reliability of different types of GPS receivers. *Sensors*, *20*(22), 6498.

Sandler, M. a. H., Andrew and Zhu, Menglong and Zhmoginov, Andrey and Chen, Liang-Chieh. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. *InProceedings of the IEEE conference on computer vision and pattern recognition 2018 (pp. 4510-4520)*.

Särkkä, S. (2008). Unscented rauch--tung--striebel smoother. *IEEE transactions on automatic control*, *53*(3), 845-849.

Särkkä, S., & Svensson, L. (2023). *Bayesian filtering and smoothing* (Vol. 17). Cambridge university press.

Spencer Jr, B. F., Hoskere, V., & Narazaki, Y. (2019). Advances in computer vision-based civil infrastructure inspection and monitoring. *Engineering*, *5*(2), 199-222.

Sun, P., Bachilo, S. M., Lin, C. W., Nagarajaiah, S., & Weisman, R. B. (2019). Dual-layer nanotube-based smart skin for enhanced noncontact strain sensing. *Structural Control and Health Monitoring*, *26*(1), e2279.

Sun, P., Draughon, G., Hou, R., & Lynch, J. P. (2022). Automated Human Use Mapping of Social Infrastructure by Deep Learning Methods Applied to Smart City Camera Systems. *Journal of Computing in Civil Engineering*, *36*(3), 04022011.

Van Overschee, P., & De Moor, B. (2012). *Subspace identification for linear systems: Theory—Implementation—Applications*. Springer Science & Business Media.

Wang, J., & Olson, E. (2016). AprilTag 2: Efficient and robust fiducial detection. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),

Welch, G., & Bishop, G. (1995). An introduction to the Kalman filter.

Xu, Y., Chen, D.-M., & Zhu, W. (2019). Operational modal analysis using lifted continuously scanning laser Doppler vibrometer measurements and its application to baseline-free structural damage identification. *Journal of Vibration and Control*, *25*(7), 1341-1364.

Yang, Y., Jung, H. K., Dorn, C., Park, G., Farrar, C., & Mascareñas, D. (2019). Estimation of full-field dynamic strains from digital video measurements of output-only beam structures by video motion processing and modal superposition. *Structural Control and Health Monitoring*, *26*(10), e2408.

Zeng, R., Wen, Y., Zhao, W., & Liu, Y.-J. (2020). View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, *6*, 225-245.

Zhang, D., Guo, J., Lei, X., & Zhu, C. (2016). A high-speed vision-based sensor for dynamic vibration analysis using fast motion extraction algorithms. *Sensors*, *16*(4), 572.

Zheng, W., Dan, D., Cheng, W., & Xia, Y. (2019). Real-time dynamic displacement monitoring with double integration of acceleration based on recursive least squares method. *Measurement*, *141*, 460-471.

Zona, A. (2020). Vision-based vibration monitoring of structures and infrastructures: An overview of recent applications. *Infrastructures*, *6*(1), 4.