Automation in Construction 140 (2022) 104376

Contents lists available at ScienceDirect

Automation in Construction

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

UAV-based bridge geometric shape measurement using automatic bridge component detection and distributed multi-view reconstruction

Yan Xu, Jian Zhang [*]

Jiangsu Key Laboratory of Engineering Mechanics, Department of Engineering Mechanics, Southeast University, Nanjing 210096, China

ARTICLE INFO

Keywords:
Bridge shape
UAV images
3D reconstruction
Distributed multi-view stereo
Object detection

ABSTRACT

UAV cameras combined with image-based reconstruction are promising non-contact tools for bridge shape inspection while computational efficiency is low, particularly on large-scale image sets. This paper describes an efficient image-based reconstruction pipeline for bridge geometry measurements. A full bridge three-dimensional reconstruction task is decomposed into multiple distributed tasks of reconstructing bridge sub-models using sub-image sets. A deep learning-based object detection method combined with Euclidean clustering is proposed to automatically generate cropped sub-image sets. An image mask generation technique based on 3D-to-2D mesh projection is used to keep valid pixels for camera pose estimation. The proposed method was validated in a laboratory test for arch shape measurement and in a field test for suspension cable shape measurement. The results show significant improvement about computational efficiency without accuracy reduction, compared with conventional image reconstruction methods.

1. Introduction

The geometric shape measurement of bridges is an important factor for bridge inspections and assessments. Monitoring the geometric profile of a bridge during various construction stages provides direct feedback regarding the as-built bridge condition to the construction control system. For aging in-service bridges, precise information on bridge deflection under static load tests provides an effective method to calibrate the remaining load-carrying capacity. In addition, horizontal and vertical clearances are important geometric parameters that must be measured with a high level of accuracy during routine bridge inspection.

Conventional sensing techniques for bridge geometric shapes include surveying tools, such as total stations and levelling. These sensors measure the spatial coordinates of sparse key points individually to generate a point-connected shape of the bridge profile and multiple measurement stations are usually configured to ensure full coverage of the entire bridge. Consequently, acquiring a high-resolution geometric shape that requires the configuration of multiple measurement stations and sensing positions is time consuming.

Computer vision techniques combined with ground-based and UAV cameras have proven to be promising non-contact tools for bridge monitoring and condition inspection [1]. Many studies have focused on developing accurate and efficient vision-based approaches for structural displacement monitoring. They have been validated to be effective for the measurement of bridge vibration properties [2], cable tension [3], and traffic-induced deflection and load rating [4]. These applications usually implement one or two fixed cameras to record image frames covering one or multiple key regions (e.g., the mid-span region) for bridges and object tracking techniques have been developed in computer vision to estimate the motions in key regions relative to an initial state, such as the beginning of video records. Displacement monitoring involves the point coordinate difference with respect to time for a specific bridge region, whereas geometric shape inspection requires acquiring the spatial coordinates of multiple bridge regions or the entire bridge. Therefore, machine vision methods for structural displacement monitoring are unsuitable for bridge geometric shape measurement.

1.1. Related work

In the literature, earlier works implemented ground-based camera systems for image acquisition and recovered geometric dimensions from a limited number of images based on features of vanishing points or stereo vision. González-Aguilera et al. [5] developed a bridge geometry extraction approach from a single oblique image using vanishing-point-based camera pose calibration. Straight lines in the bridge geometry were extracted to compute the vanishing points and recover the

[*] Corresponding author.
E-mail addresses: xuyan0909@live.com (Y. Xu), jian@seu.edu.cn (J. Zhang).

https://doi.org/10.1016/j.autcon.2022.104376
Received 11 February 2022; Received in revised form 16 May 2022; Accepted 19 May 2022
Available online 27 May 2022

projection matrix between the image plane and bridge coordinate system. Because the proposed pipeline requires a priori of straight lines in the bridge profile, it is unsuitable for bridges with complex shapes (e.g., horizontally curved ones). In addition, the resolution and accuracy of dimension measurements from a single image decrease with a larger field-of-view, and thus, using single images is inappropriate for the measurement of large-span bridges. Riveiro et al. [6] proposed a non-contact measurement approach for bridge minimum vertical under-clearance. A camera network with proper overlapping views was utilised for image acquisition and the multi-view images were processed based on stereo vision algorithms to reconstruct three-dimensional (3D) coordinates of key bridge positions located on the lower beam and pavement surfaces. Consequently, this approach only recovers 3D coordinates for the sparse points of interest.

As an emerging technology for consumer-grade UAVs in recent years, the implementation of numerous overlapping aerial photographs for 3D bridge modelling has received significant attention as a low-cost option for 3D imaging. Image-based 3D reconstruction is used to create 3D point clouds of a bridge structure from a collection of multi-view overlapping images, providing a potential solution for bridge geometric sensing. Compared with terrestrial laser scanning, image-based 3D reconstruction is an inexpensive and efficient method for 3D mapping [7,8], however, the achievable accuracy is low [9]. Image-based 3D reconstruction software packages are widely available for users including open-source (VisualSfM [10], COLMAP [11] and OpenMVS [12]) and commercial (Agisoft Metashape [13] and DJI Terra [14]) packages. The main framework consists of two steps: structure from motion (SFM) to estimate the camera poses of each image and multi-view stereo (MVS) reconstruction to recover a dense model by fusing the depth estimation of each image. A classical SFM pipeline consists of three major steps: feature extraction, feature matching and parameter solving based on iterative bundle adjustment [15,16]. There are some review studies providing photogrammetric 3D reconstruction techniques for civil infrastructure [15,17].

The reconstructed point clouds obtained from image-based 3D reconstruction can be applied to geometric dimension measurements and surface condition assessments. Golparvar-Fard et al. [18] developed an automatic SFM-based reconstruction method to verify the daily as-built project status of construction sites. Kassotakis et al. [19] implemented SFM photogrammetry techniques to quantify the effect of geometric uncertainty on the structural behaviour of arches under laboratory conditions. Moon et al. [20] proposed a method for generating and merging hybrid point cloud data acquired from laser scanning and UAV-based image processing for earthwork projects. Morgenthala et al. [21] presented a framework for automated unmanned-aircraft-system-based inspections of bridges for condition assessment and discussed flight path planning, structural surface model reconstruction and surface defect detection schemes. Chen et al. [22] provided a case study of UAV bridge inspection using 3D reconstruction and discussed the quality evaluation mechanism for a point cloud model. Jiang and Bai [8] presented a 3D reconstruction method for elevation determination at construction sites using drone-based, low-high orthoimage pairs. The image-derived point cloud can be further integrated with the point clouds from other sensors for specific applications. For example, thermal images were used to obtain the thermal transmittance values of external walls [23]. Bacharidis et al. [24] proposed a multimodal fusing scheme of stereoscopic images to reconstruct 3D models of a structure's facade that integrates a two-dimensional (2D) building skeleton from viewed scenes with a depth information layout extracted from a stereoscopic layout.

Image-based point clouds were further implemented for the as-built modelling of structural elements. Pan et al. [25] presented a semi-automated framework for generating structural surface models of heritage bridges. A top-down method for bridge component classification was applied to label the SFM-based point clouds and each element was independently applied for surface modelling. Hu et al. [26] developed a structure-aware learning-based cable-stayed bridge 3D reconstruction framework. An encoder-decoder model was proposed to import both multi-view images and a dense point cloud model using the SFM to predict high-level structural relation graphs and low-level 3D geometric shapes. The method was validated in two actual cable-stayed bridges and the acquired model was found to be suitable for inspection path planning. Xu et al. [27] proposed an object-level volumetric reconstruction method for as-built buildings based on a two-stage region-based end-to-end 3D object detection network. To build a primitive of a symmetric cuboid, building objects (e.g., walls, windows, and doors) were automatically detected and reconstructed using an end-to-end solution.

For geometry quantification, measurement accuracy, efficiency and point cloud completeness [28] are common metrics to evaluate working performance of image-based 3D reconstruction. The measurement accuracy is highly dependent on the SFM process that computes camera poses and a sparse point model. The accuracy is relatively unsatisfied compared with terrestrial laser scanners [18], especially for large-scale scenarios. Acquiring closer-up images is effective to grasp more structural local details and improve feature matching process. But it is at the cost of increased image number and higher computation efforts. To overcome the length scale issue for large-scale structure reconstruction, Khaloo and Lattanzi [29] developed a hierarchical dense SFM approach for bridge profile reconstruction. This technique first generates a series of dense point clouds at varying length scales using the SFM and dense pixel-wise image matching, and then merges these individual, multiscale point clouds into a global point cloud model. This multi-scale photogrammetry technique was also applied to create high-resolution 3D point clouds of the dam for the assessment of their overall quality, as well as the detection of flaws and defects [30]. Besides, Xu et al. [31] proposed a 3D building reconstruction method based on geometrical characteristics of straight lines and vanishing points, and achieved similar results about camera pose compared with the traditional SFM results.

Removing unwanted features in images offers a few advantages, including decreasing computational time and improving the quality of the region of interest [32]. Saovana et al. [33] proposed an unwanted-feature removal system for images of repetitive infrastructures such as piers. A deep learning model was trained on real-world infrastructure images to aid the SFM process by removing pixels belonging to unwanted classes. The authors [34] also developed a point cloud classification framework by projecting image-based instance segmentation results to 3D SFM point clouds for bridge component modelling. Narazaki et al. [35] investigated vision-based automated bridge component recognition techniques for visual inspection of bridges after earthquakes. It consists of 10-class scene classification and 5-class bridge component classification. A semantic segmentation model with sequential configuration of scene and component classification tasks is demonstrated to be effective. Mirzazade et al. [36] studied the bridge component detection method using deep learning-based semantic segmentation. Two encoder-decoder architectures (i.e. SegNet and U-Net) were implemented for background removal, and SegNet showed better performance in terms of prediction accuracy and computation efficiency.

1.2. Scope of this study

Although a few existing studies have reported the implementation of image-based 3D reconstruction techniques for dimension measurements, there are some limitations in terms of measurement accuracy and efficiency. First, the SFM process computes camera poses based on local feature correspondences in multi-view images. Bridge structures contain multiple highly similar components (e.g., piers and cable clamps) along the length direction which might cause a high proportion of incorrect feature matches [15]. In addition, bridge structures captured in images are usually slim, occupying a small portion of the entire image. The

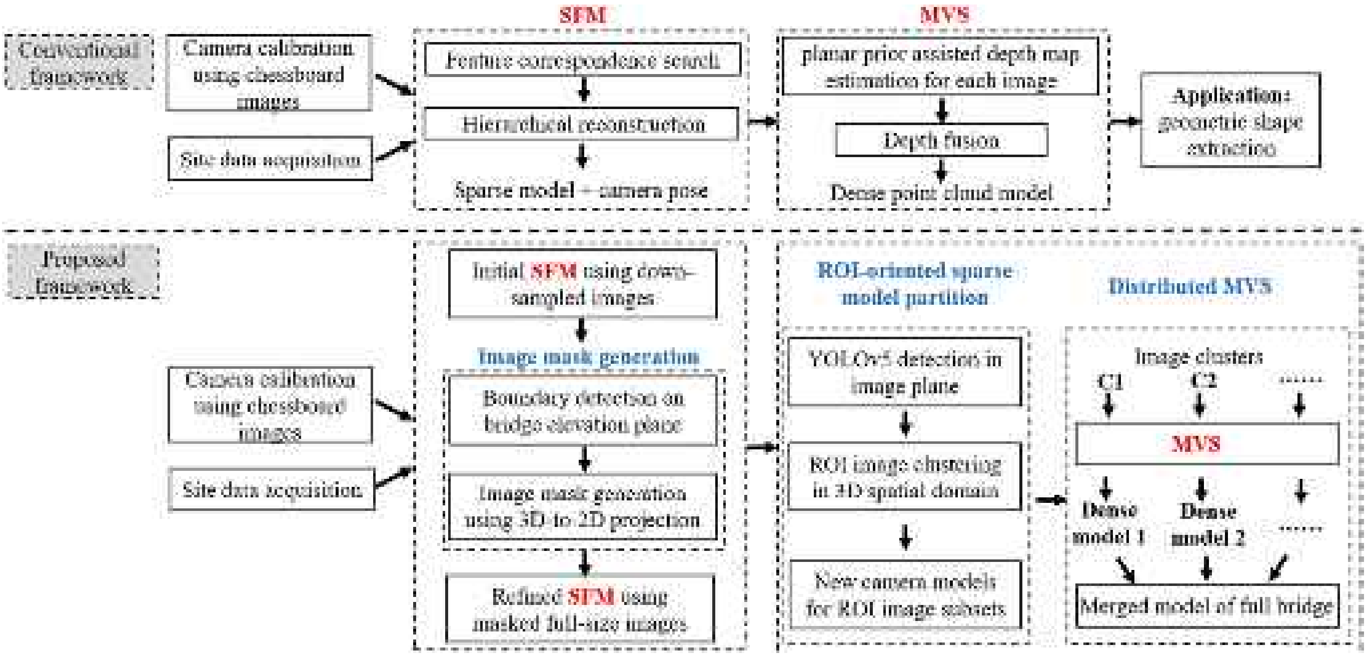

Conventional framework
Camera calibration using chessboard images
Site data acquisition
SFM
Feature correspondence search
Hierarchical reconstruction
Sparse model + camera pose
MVS
planar prior assisted depth map estimation for each image
Depth fusion
Dense point cloud model
Application: geometric shape extraction
Proposed framework
Camera calibration using chessboard images
Site data acquisition
Initial SFM using down-sampled images
Image mask generation
Boundary detection on bridge elevation plane
Image mask generation using 3D-to-2D projection
Refined SFM using masked full-size images
ROI-oriented sparse model partition
YOLOv5 detection in image plane
ROI image clustering in 3D spatial domain
New camera models for ROI image subsets
Distributed MVS
Image clusters
C1
C2
......
MVS
Dense model 1
Dense model 2
......
Merged model of full bridge

Fig. 1. Framework of image-based 3D reconstruction for bridge shape measurement proposed in this study.

feature matches among background pixels (e.g., sky and river) might deteriorate the estimation results of the camera poses. Second, the MVS process for the computation of a dense point cloud is usually applied to each individual full-image and requires high computational effort and memory storage. Background pixels in images provide no useful information but are taken into consideration during the dense reconstruction process.

In this study, a new image-based 3D reconstruction pipeline for accurate and efficient geometry measurements of bridge structures is proposed. The main contributions of this study are as follows. First, to overcome the poor performance of SFM owning to false feature matches in repetitive bridge components and background pixels, sparse reconstruction is refined by image masking based on 3D-to-2D mesh projection for removing background pixels. Second, the dense reconstruction step, which usually requires high computational effort, is converted into multiple independent tasks for reconstructing bridge sub-regions in a unified coordinate system. By considering the repetitive occurrence of some key bridge components (e.g., piers and cable clamps), the full bridge model can be decomposed into multiple regions of interest (ROIs) along the length direction centred at these key components. For each ROI, sub-image clusters were then automatically generated based on a deep learning-based object detection algorithm. A plane prior assisted MVS algorithm is used for dense reconstruction to improve point cloud completeness.

Section 2 describes the primary methodologies used in the proposed image-based 3D reconstruction pipeline. Section 3 describes the validation of the proposed method in a laboratory bridge structure for arch shape measurement. Section 4 presents an application example of a suspension bridge structure for cable shape measurement. Conclusions are presented in Section 5.

2. Methodologies

The conventional image-based 3D reconstruction framework [16] is presented in Section 2.1, and the proposed improvements in the SFM and MVS processes are clarified in Sections 2.2 and 2.3.

2.1. Proposed 3D image reconstruction framework using SFM and MVS

SFM algorithms enable the automatic estimation of camera poses and construction of a sparse point cloud from multi-view overlapped images. Based on the outputs of the SFM, the MVS performs image depth computation and fusion to generate a dense point cloud model.

The SFM process used in this study, following the steps of feature correspondence search among image pairs and hierarchical image reconstruction [11] is shown in Fig. 1. In the first stage, feature extraction is executed to detect key-points represented by descriptor vectors using DSP-SIFT [37]. For feature matching, the Euclidean distance between feature descriptors was used as metric to discover similar feature pairs among images. With the geo-information extracted from image EXIF data, the match search is constrained to its nearest spatial neighbours. Geometric verification is applied to check the correctness of image pairs by using projection geometry constraints and to remove feature correspondence outliers using the robust estimation technique RANSAC [38]. In the second stage, the images are partitioned into multiple clusters with overlaps for parallel computation. For each partition, an initial two-view reconstruction was conducted using an image pair with a sufficiently overlapped region to extract their poses and 3D points. New images overlapping with the current reconstructed model were registered iteratively by solving the perspective-n-point problem [39], and new 3D points were added to the model by image triangulation. To eliminate accumulated errors, bundle adjustment was performed to refine the camera poses and 3D points by minimising the total re-projection error between the point projections and their corresponding feature points. Finally, the overlapped sub-models were merged into a single reconstruction model, which was further refined by two rounds of point triangulation and bundle adjustment over the entire scene data.

Given the estimated camera calibration parameters from the SFM process, the MVS computes the depth map for each image based on epipolar geometry, and fuses the depth estimates into a globally consistent point cloud of the entire scene. Among the multiple MVS algorithms available, planar prior assisted patch-match multi-view stereo [40] was used in this study to provide robust depth estimation in low-textured areas. It consists of three stages: generating sparse initial depth estimates via multi-view aggregated photometric consistency

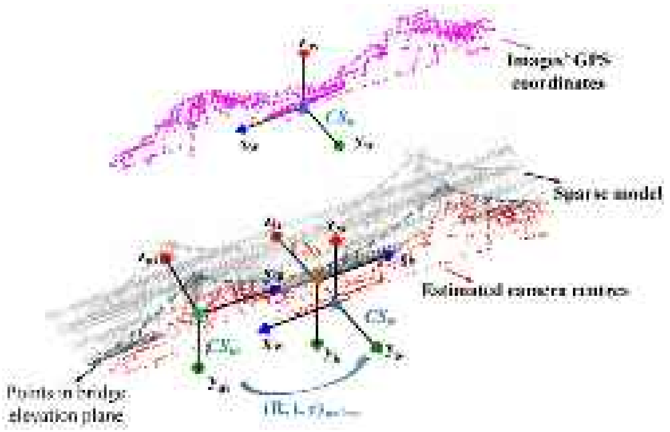


Fig. 2. Demonstration of rigid geometric alignment between the world coordinate system (CS_w) and the measured coordinate system (CS_m) of the sparse model.

cost, triangulating credible correspondences to generate planar models, and generating plane-awareness depth maps by joint consideration of photometric consistency and planar priors. By incorporating the joint matching cost, this method is suitable for both planar and non-planar regions. Finally, depth fusion filters out outliers and compacts the point cloud model from the consistency evaluation. The depth map for each image is transformed into a unified coordinate system to obtain 3D points, and then it is further converted into the 2D neighbouring image planes to derive the corresponding projection of the depth and image coordinates. Outliers are removed by applying thresholds on the depth and image coordinate differences among neighbouring images. Consistent 3D points are averaged for the final point cloud model generation. The dense point cloud model of the bridge is scaled using known geometric dimensions for the geometric shape measurement.

To improve measurement accuracy and computational efficiency, the proposed framework for bridge shape measurement is illustrated in Fig. 1. In sparse reconstruction, the framework incorporates a preprocessing step of image mask generation based on the 3D-to-2D mesh projection to ensure feature correspondence searches among only the valid structural regions. Besides, an additional module, an ROI-oriented sparse model partition, was added before dense reconstruction. An object detection framework, YOLOv5 [41], was implemented to identify the ROIs in each image and to assist in partitioning the sparse model into multiple spatially separated sub-models and image sets for distributed dense reconstruction. A custom-made package realising the proposed method was developed in Python, partially referring to the existing modules in COLMAP [11,42]. An detailed introduction to image mask generation and ROI-oriented sparse model partitioning is given in the next two subsections.

2.2. Image mask generation

To ensure a feature correspondence search only in valid bridge regions, image masks were created to exclude background pixels in the SFM. Instead of directly extracting pixels belonging to the bridge from the 2D image representation, an initial SFM for 3D environment modelling was constructed and the valid region in the 3D spatial domain was selected. This is then projected onto the 2D image planes for mask generation.

An initial SFM was conducted using multi-view images to build a coarse incomplete sparse model and derive their camera poses. Because the requirement for estimation accuracy is low, the settings were adjusted to improve efficiency: the images were down-sampled by a factor of 4 times, SIFT was used as the feature descriptor, and the image number for hierarchical sub-models was 100. Here, the SFM calculation used the existing modules in COLMAP. The estimated camera poses and sparse model were unscaled and in an unknown coordinate system denoted as CS_m.

To automatically extract the valid region in 3D space, the sparse model was first rigidly aligned with the world coordinate system (CS_w) (shown in Fig. 2), and the points within a threshold distance from the closest camera centres were retained for plane fitting on the bridge elevation plane. Coplanar points in the fitted elevation plane were converted into a single polygon shape representing valid regions. The GPS coordinates extracted from the image EXIF data are in the CS_w coordinate system. The estimated camera centre coordinates and the images' GPS coordinates are denoted as $\mathbf{P} = \{\mathbf{p_1}, \mathbf{p_2}, \dots, \mathbf{p_n}\}$ and $\mathbf{Q} = \{\mathbf{q_1}, \mathbf{q_2}, \dots, \mathbf{q_n}\}$. The two coordinate sets were deducted by the coordinates of centroids ($\overline{\mathbf{P}}$ and $\overline{\mathbf{Q}}$) individually to derive the centred coordinates $\mathbf{P_c}$ and $\mathbf{Q_c}$. The alignment task is equivalent to finding a rotation matrix $\mathbf{R}_{m->w}$, a translation vector $\mathbf{t}_{m->w}$, and a scale factor $c_{m->w}$ such that

$$(\mathbf{R}, \mathbf{t}, c)_{m->w} = argmin \sum_{i=1}^{n} \left\| c_{m->w} \mathbf{R}_{m->w} \mathbf{p_i} + \mathbf{t}_{m->w} - \mathbf{q_i} \right\|^2. \tag{1}$$

The optimal rotation matrix $\mathbf{R}_{m->w}$ was computed using the singular value decomposition on the covariance matrix $\mathbf{S} = \mathbf{P_c}\mathbf{Q_c}^T$ of the two centred coordinate sets:

$$\mathbf{S} = \mathbf{U\Sigma V}^T, \mathbf{R}_{m->w} = \mathbf{VU}^T. \tag{2}$$

The optimal scale factor was computed by the division of two matrix traces,

$$c_{m->w} = tr(\mathbf{\Sigma}) / tr\left(\mathbf{P_c}\mathbf{P_c}^T\right), \tag{3}$$

and the optimal translation was computed as the centroid projection difference,

$$\mathbf{t}_{m->w} = \overline{\mathbf{Q}} - c_{m->w}\mathbf{R}_{m->w}\overline{\mathbf{P}}. \tag{4}$$

The sparse model and camera coordinate estimates in the CS_m were transformed to the CS_w. The nearest neighbour search was used to identify the nearest camera coordinates for each 3D point. A distance threshold D (e.g., 20 m on site) was set on the point pair distance to filter out background regions. The RANSAC plane fitting was conducted to find the dominant plane in inlier points that is equivalent to the bridge elevation plane. The alpha shape algorithm [43] was utilised to compute the approximate boundaries of coplanar points, which were then unified into a single polygon shape. For a planar sparse model of one bridge elevation plane, different alpha values were applied to generate the candidate polygon shapes with the results shown in Fig. 3. By considering a whole coverage of merely suspension cable regions, the alpha value of 0.7 was taken for the calculation. The polygon vertices are denoted as $\mathbf{V}^w = \{\mathbf{v}_1^w, \mathbf{v}_2^w, \cdots, \mathbf{v}_k^w\}$ in theCS_w. The vertices' coordinates were transformed to theCS_m as $\mathbf{V}^m = \{\mathbf{v}_1^m, \mathbf{v}_2^m, \cdots, \mathbf{v}_k^m\}$ by inversely applying $(\mathbf{R}, \mathbf{t}, c)_{m->w}$.

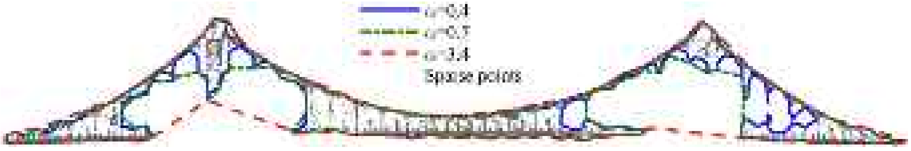


Fig. 3. Demonstration of polygon shapes generated from a planar sparse model by alpha-shape algorithm.

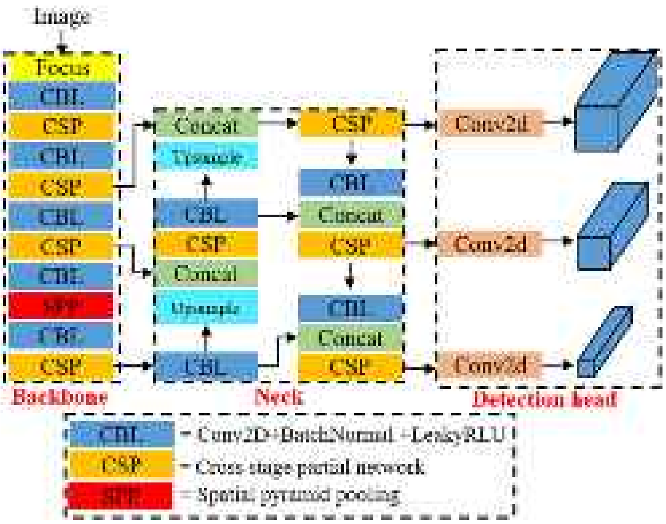


Fig. 4. YOLOv5 architecture.

Given the camera models for each image, the polygon shape in the CS_m can be easily projected onto the 2D image plane, following the pinhole camera model,

$$\alpha\begin{bmatrix}\mathbf{u}\\1\end{bmatrix}=\mathbf{K}[\mathbf{R}|\mathbf{t}]_{\mathbf{I}}\begin{bmatrix}\mathbf{V}^m\\1\end{bmatrix}. \tag{5}$$

where **K** is camera intrinsic matrix pre-computed using chessboard images, which is fixed for all images, $[\mathbf{R}|\mathbf{t}]_{\mathbf{I}}$ is the camera orientation and position with respect to the CS_m, which is specific to each image outputted in the initial SFM process, $\begin{bmatrix}\mathbf{V}^m\\1\end{bmatrix}$ is the homogenous coordinates of polygon vertices in the CS_m, **u** is the 2D image coordinates of polygon vertex projection, which is to be computed, and α is a scale factor. For each image, the projection of the polygon shape in the 2D image plane was derived following the above equation, and the intersection part with the image boundary was taken as the image mask. To ensure a whole coverage of valid regions, the mask image was dilated using a disk-shaped structuring element of 20 pixel radius. Using the full-sized masked images, a second round of the hierarchical SFM was conducted to derive an accurate estimation of camera poses and sparse model.

2.3. ROI-oriented sparse model partition

The module of ROI-oriented sparse model partition was used for the realisation of distributed dense reconstruction. It runs multiple independent tasks of dense reconstruction of partial regions using a small group of images. The process is as follows: ROIs, such as cable clamps for cable shape measurement, are detected from images using a deep learning-based object detection framework, YOLOv5. Because the detected ROIs are unordered in the 2D image plane, the ROI centroids are re-projected onto the bridge elevation plane in the 3D spatial domain for clustering, and the detected ROI images are labelled with cluster indices. This partitions the sparse model spatially into multiple submodels oriented at the ROIs with consistently cropped ROI image sets.

2.3.1. ROI detection in 2D image plane

A deep learning-based object detection framework, YOLOv5 was selected to identify the ROI regions in the 2D image plane. The YOLO family is an efficient one-stage detection algorithm that directly obtains the location and category of each object class via regression in real time. YOLOv3 [44] is a popular choice for real-time object detection tasks that uses independent logistic classifiers instead of the softmax function to determine the class of an input image. YOLOv4 [45] made improvements using some state-of-the-art methods to make the corresponding modification. It incorporates a cross-stage partial structure and a spatial pyramid pooling block as the backbone to enhance the learning capacity and increase the receptive field. YOLOv5 [41], released in 2020 is similar to YOLOv4. It is completely compiled by PyTorch, and uses autolearning bounding box anchors to adjust and optimise the choice of anchors. The model is composed of three main components: a backbone, neck and detection head. The YOLOv5 architecture is shown in Fig. 4.

The backbone incorporates Focus structure and cross stage partial networks (CPSNet) to aggregate and form image key features. The input image with 640 × 640 × 3 resolution goes through the Focus structure first, using the slicing operation and a convolution operation to a 320 × 320 × 32 feature map. The CBL module is a basic convolution module representing Conv2D + BatchNormal +LeakyRLU. The BottleneckCSP block mainly performs feature extraction on the feature map, extracting rich information from the image, and effectively reduces network parameters by reducing gradient information duplication. The spatial pyramid pooling (SPP) module mainly increases the receptive field of the network, and acquires features of different scales. The neck consists of a normal feature pyramid network (FPN) and path aggregation network (PANet). The FPN layer conducts a top-down pathway to extract strong semantic features, and the feature pyramid conveys robust positioning features from the bottom up. Through combining image features in a series of network layers, it improves the object identification in various sizes and scales. The head is responsible for final detection with the outputs of class possibility and bounding boxes.

A field test study was conducted for bridge cable shape measurement, and the ROIs were about cable clamps. A cable clamp dataset was created using images of the same test bridge taken under various lighting conditions and camera-to-bridge distances. The total image number of raw images was 215, and the image resolution was 5466 × 3630 pixels. The cable clamp number in a single image varied from 2 to

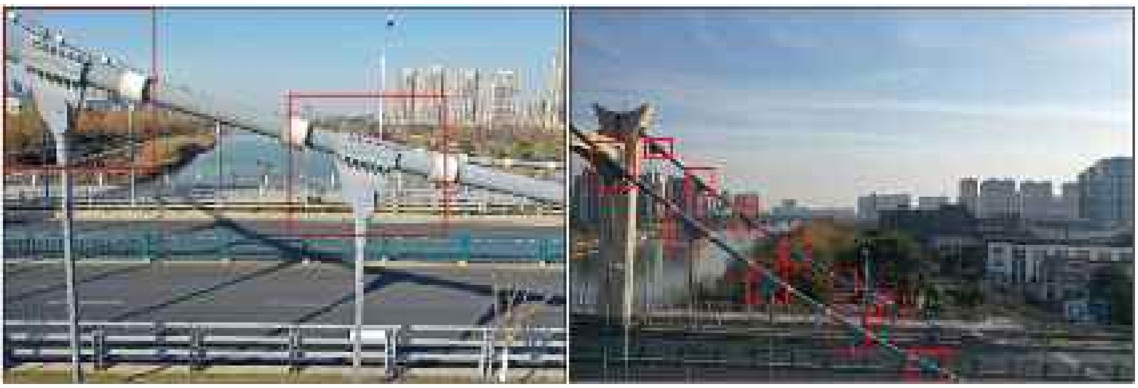


Fig. 5. Object detection results for cable clamps in two sampled images.

23 with an average of 9.4. The raw images were down-sampled by a factor of 2, and then tiled to sub-images with dimensions of 640 × 640 pixels, leading to a dataset of 864 images with annotations of cable clamps.

The network was trained using the PyTorch framework. The training computer consists of an i9-10900k CPU, an NVIDIA RTX3090 GPU, and 32GB RAM. Images were divided into training and test images in the split ratio of 8:2. For the training, the batch size and epoch number were set to 16 and 120, respectively. The SGD optimiser was used with an initial and final learning rates of 0.01 and 10^{-5}. After training, the mean average precision at the intersection over union of 0.5 reached 97.2%.

The trained model was implemented to detect the cable clamp regions in multi-view UAV images. The raw images were down-sampled by a factor of 2 and tiled to sub-images with a size of 640 × 640 pixels before object detection. After detection, the extracted bounding boxes were recovered to the raw image scale. Fig. 5 presents the object detection results for two sample images captured under different lighting conditions and camera-to-bridge distances. The trained model successfully detected ROIs with rectangle sizes varying from 222 × 224 to 1470 × 1560 pixels. The detected regions were compared with the image masks derived in Section 2.2, and only those with an overlap ratio of >50% were considered valid. Valid regions were segmented into cropped ROI images for further analysis.

2.3.2. ROI image clustering in 3D spatial domain

The previous step extracted the ROI positions in the 2D image plane, whereas the spatial information of these ROIs in the bridge structure was unknown. Therefore, ROI image clustering was conducted by back-projection from the 2D pixel coordinates to the 3D point coordinates based on the estimated camera model in the SFM. Assuming that the back-projection of the ROI centroids lies on the bridge elevation plane, it can be simplified as a 2D-to-2D transformation from the 2D image plane to the 2D bridge elevation plane.

Following the description in Section 2.2, the perspective transformation from the 2D image plane to the 3D coordinate system CS_m of the sparse model is given in Eq. (5). The transformation $(\mathbf{R}, \mathbf{t}, c)_{m->w}$ from the CS_m to the CS_w and the plane model for the bridge elevation were precomputed. A new bridge coordinate system CS_b was created with the XY plane consistent with the bridge elevation. The rigid transformation from the CS_w to the CS_b is denoted as $(\mathbf{R}, \mathbf{t})_{w->b}$. The rotation matrix was composed of three unit vectors, $\boldsymbol{R}_{w->b} = \begin{bmatrix} \overrightarrow{x}_b \\ \overrightarrow{y}_b \\ \overrightarrow{z}_b \end{bmatrix}$, where the z-axis vector $\overrightarrow{\mathbf{z}}_{\mathbf{b}}$ is equivalent to the normal vector of the elevation plane, the x-axis vector $\overrightarrow{\mathbf{x}}_{\mathbf{b}}$ is the projection of the x-axis vector $\overrightarrow{\mathbf{x}}_{\mathbf{w}} = [1 \;\; 0 \;\; 0]$ in CS_w onto the elevation plane with normalisation that is $\overrightarrow{\mathbf{x}}_{\mathbf{b}} = \overrightarrow{\mathbf{x}}_{\mathbf{w}} - \text{dot}\left(\overrightarrow{\mathbf{z}}_{\mathbf{b}}, \overrightarrow{\mathbf{x}}_{\mathbf{w}}\right) \cdot \overrightarrow{\mathbf{x}}_{\mathbf{w}} / norm\left(\overrightarrow{\mathbf{x}}_{\mathbf{w}} - \text{dot}\left(\overrightarrow{\mathbf{z}}_{\mathbf{b}}, \overrightarrow{\mathbf{x}}_{\mathbf{w}}\right) \cdot \overrightarrow{\mathbf{x}}_{\mathbf{w}}\right)$, and the y-axis vector $\overrightarrow{\mathbf{y}}_{\mathbf{b}}$ is the cross product between $\overrightarrow{\mathbf{z}}_{\mathbf{b}}$ and $\overrightarrow{\mathbf{x}}_{\mathbf{b}}$. The translation vector $\mathbf{t}_{w->b}$ is equivalent to $-\mathbf{R}_{w->b}\overline{\mathbf{V}}^{w}$, where $\overline{\mathbf{V}}^{w}$ is centroids of co-planar points in the elevation plane.

For an image point $\mathbf{u_i} = \begin{bmatrix} u_i \\ v_i \end{bmatrix}$ on a cropped ROI centroid in the full-image view, the perspective projection from the 2D image plane to the bridge local coordinate system CS_b can be summarised as

$$\begin{aligned} \alpha \boldsymbol{u}_i &= \mathbf{K}[\mathbf{R}|\mathbf{t}]_{\mathbf{I}} \boldsymbol{V}_i^m \\ \boldsymbol{V}_i^w &= c_{m->w} \boldsymbol{R}_{m->w} \boldsymbol{V}_i^m + \boldsymbol{t}_{m->w} \\ \boldsymbol{V}_i^b &= \boldsymbol{R}_{w->b} \boldsymbol{V}_i^w + \boldsymbol{t}_{w->b} \end{aligned} \tag{6}$$

where $\boldsymbol{V}_i^m$, $\boldsymbol{V}_i^w$ and $\boldsymbol{V}_i^b$ correspond to the back-projection points of $\mathbf{u_i}$ in CS_m, CS_w and CS_b, respectively. The previous projection equation is compacted as

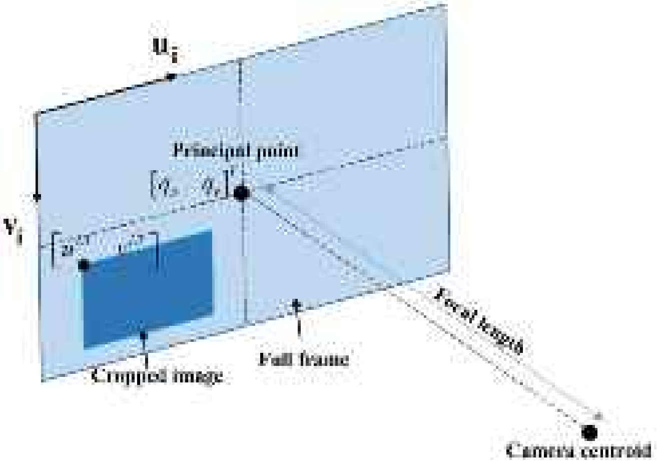


Fig. 6. Demonstration of the new camera model for an cropped image.

$$\begin{aligned} \alpha \begin{bmatrix} \boldsymbol{u}_i \\ 1 \end{bmatrix} &= \boldsymbol{M}_{3\times4} \begin{bmatrix} \boldsymbol{V}_i^b \\ 1 \end{bmatrix} \\ \boldsymbol{M}_{3\times4} = \boldsymbol{K}[\boldsymbol{R}|\boldsymbol{t}]_{\mathrm{I}} &\begin{bmatrix} c_{m->w}\boldsymbol{R}_{m->w} & \boldsymbol{t}_{m->w} \\ \boldsymbol{O}_{1\times3} & 1 \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{R}_{w->b} & \boldsymbol{t}_{w->b} \\ \boldsymbol{O}_{1\times3} & 1 \end{bmatrix}^{-1} \end{aligned} \tag{7}$$

where $\mathbf{M}_{3\times4}$ is a 3D homography matrix with 3 × 4 dimension directly from the 2D image plane to the CS_b. Because the back-projected point $\boldsymbol{V}_i^b = \begin{bmatrix} x_i^b \\ y_i^b \\ z_i^b \end{bmatrix}$ lies on the XY plane with $z_i^b = 0$, the projection model by neglecting the coordinate z_i^b can be as downgraded to a 2D-to-2D projection model following

$$\alpha \begin{bmatrix} u_i \\ v_i \\ 1 \end{bmatrix} = \begin{bmatrix} m_{11} & m_{12} & m_{14} \\ m_{21} & m_{22} & m_{24} \\ m_{31} & m_{32} & m_{34} \end{bmatrix}_{3\times3} \begin{bmatrix} x_i^b \\ y_i^b \\ 1 \end{bmatrix} \tag{8}$$

where m_{jk} denotes the matrix element at row j and column k in the 3D homography matrix $\mathbf{M}_{3\times4}$.

The 2D pixel coordinates $\mathbf{u_i}$ of ROI centroids were all back-projected using Eq. (8) to derive the 3D coordinates $\boldsymbol{V}_i^b$ (with the Z values filled by 0) of ROI centroids in a unified coordinate system CS_b. The hierarchical clustering with the measure metric of Euclidean distance was conducted to separate the ROI centroids $\boldsymbol{V}_i^b$ into groups. The cluster labels of the ROI centroids were assigned to the cropped ROI images for the separation. Multiple image subsets with varied image sizes were generated for distributed 3D reconstruction.

2.3.3. New camera models for ROI image subsets

The cropped-image subsets were implemented separately for the dense reconstruction of each ROI, generating multiple ROI-oriented point cloud models in a unified coordinate system. Because the dense reconstruction requires the input of multi-view images and their camera parameters, the camera models were corrected for each cropped image to adapt to the image changes.

For the raw image with the full size, the initial equivalent camera intrinsic matrix after lens distortion correction is

$$\mathbf{K}^{\mathbf{I}} = \begin{bmatrix} f_u & 0 & q_u \\ 0 & f_v & q_v \\ 0 & 0 & 1 \end{bmatrix} \tag{9}$$

where f_u and f_v are the focal lengths of the camera in terms of pixel

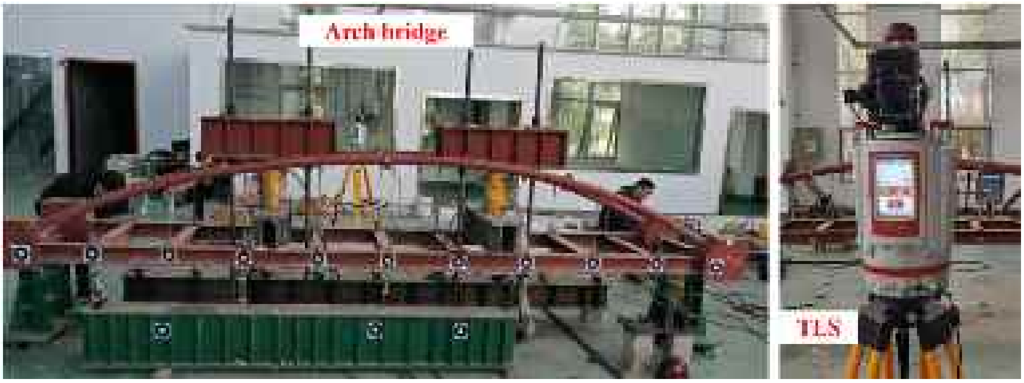


Fig. 7. Photo of the laboratory steel arch bridge.

dimensions along the image width and height directions, respectively; and $\begin{bmatrix} q_u & q_v \end{bmatrix}^T$ are the image coordinates of the principal point. A cropped image was generated from the raw image, truncating from the top-left pixel coordinates of $\begin{bmatrix} u^{LT} & v^{LT} \end{bmatrix}^T$ with the new region dimension of $\begin{bmatrix} w^N & h^N \end{bmatrix}$, as shown in Fig. 6. A virtual new camera model for the cropped image was constructed that inherits most of the initial camera parameters, including the camera extrinsic matrix and equivalent focal lengths. Only the image coordinates of the principal point were corrected to $\begin{bmatrix} q_u - u^{LT} & q_v - v^{LT} \end{bmatrix}^T$. It is acceptable to contain negative values in the principal point coordinates. The camera models were varied for each cropped image.

Given the new camera models and cropped image subsets, the ROI sub-models were computed in a distributed manner and finally merged into a single full model for bridge shape extraction.

3. Laboratory test for bridge beam shape measurement

This section discusses the proposed method onto a laboratory arch bridge structure for arch and beam shape measurements.

3.1. Test information

The tested bridge was a steel through-arch bridge with a total span length of 5.6 m, as shown in Fig. 7. The bridge deck system consisted of two primary beams and eleven transverse beams. The deck system and two arch ribs were fixed by bolt connections on the bridge ends that settled down the supporting system with a fixed-pinned constraint. Several circular target patterns were attached to one primary beam and a fixed ground beam as the ROI annotations. The arch shape and positions of the target regions on the primary beam were reconstructed from multi-view images and a terrestrial laser scanner (TLS), respectively.

A UAV (DJI Mavic 2 Pro) was used for image acquisition, and totally 59 UAV images were acquired from one side of the bridge. The acquired image resolution was 5472 × 3648 pixels. The camera intrinsic parameters were calibrated ahead of the test by observing chessboard patterns from multiple views. As a reference sensor, the TLS Reigl VZ-400i was used to record the point cloud model of the bridge (which is denoted as Model 1). The measurement accuracy and precision are 5 mm and 3 mm, respectively. The UAV images after camera distortion correction were reconstructed to 3D point cloud models using three methods (i.e., COLMAP SFM + COLMAP MVS, COLMAP SFM + Plane prior assisted MVS, and the proposed method), which are denoted as Models 2–4, respectively. During the SFM calculation, the camera intrinsic parameters remained fixed, and only the camera poses were optimised in the bundle adjustment. For each image pair, the minimum number of feature matching inliers to be considered geometrically verified was set to 15, and the max allowed re-projection error was set to 1 pixel.

3.2. Analysis results

The SFM results using the full and masked images are listed in Table 1. The results show that the sparse model using masked images reduces the number of tracked 3D points from 22,757 to 9572 by neglecting the background pixels. The mean track length, corresponding to the average number of images in which a feature is successfully tracked, increases slightly from 3.258 to 3.656. The mean re-projection error, which indicates the difference between the image projections of the calculated 3D points and their detected feature positions on images, is similar at approximately 0.364 pixels. The comparison indicates that incorporating image masks of valid bridge regions in the SFM leads to a smaller sparse model and improved computational efficiency. The improvement in accuracy is not obvious, probably because the salient features detected from the background objects in this indoor condition are distinguished and robust for matching, unlike the features in the sky or river in outdoor cases that provide little positive assistance for camera pose estimation.

The dense reconstruction results were used to measure the bridge arch and girder shapes. The four point cloud models shown in Fig. 8 (i.e., Model 1 by TLS, Model 2 by COLMAP SFM + COLMAP MVS, Model 3 by COLMAP SFM + Plane prior assisted MVS, and Model 4 by the proposed method) are given for comparison. The three image reconstruction models were scaled by the given bridge length and then aligned with Model 1 by the iterative closest point.

The computation time of the dense reconstruction for Model 2–4 is listed in the second row of Table 2. It shows that,

- By incorporating plane prior assisted MVS, the computation time used for the dense reconstruction for the same 59 full images of the arch bridge decreased from 130.78 min to 29.90 min.
- By using the proposed distributed MVS, the dense reconstruction for a single ROI took averagely 1.95 min. The total computation time is

Table 1
Reconstruction statistics for the two SFM process.

SFM results	Computation time (min)	Tracked 3D Points	Mean track length	Mean observations per image	Mean re-projection error (pixel)
Using full images	2.897	22,757	3.258	1256.75	0.3644
Using masked images	1.966	9572	3.656	593.07	0.3634

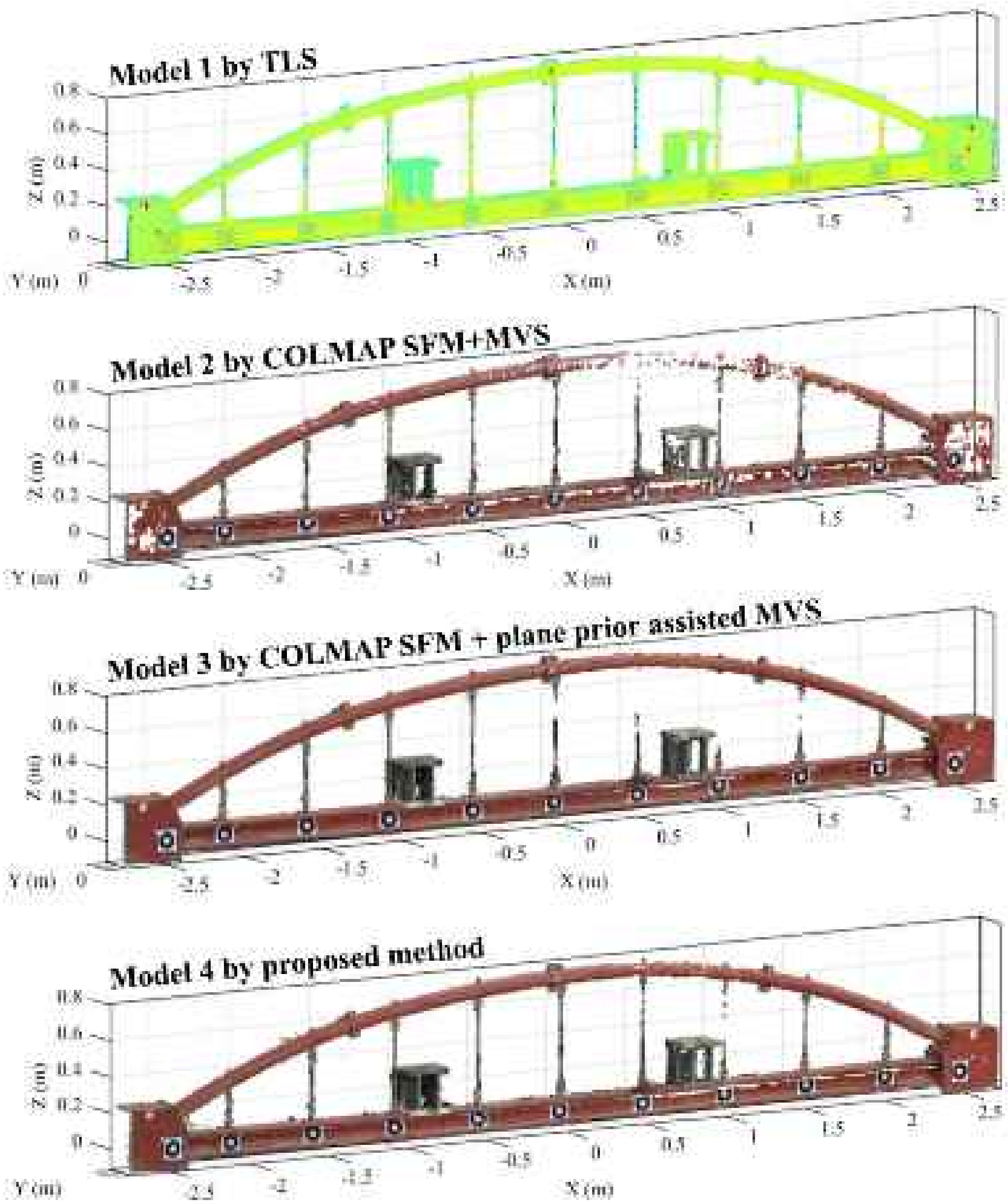

Model 1 by TLS
Model 2 by COLMAP SFM+MVS
Model 3 by COLMAP SFM + plane prior assisted MVS
Model 4 by proposed method
Z (m)
0.8
0.6
0.4
0.2
0
Y (m)
X (m)
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5

Fig. 8. Four point cloud models for comparison: Model 1 by TLS, Model 2 by COLMAP SFM + COLMAP MVS, Model 3 by COLMAP SFM + Plane prior assisted MVS, and Model 4 by the proposed method.

Table 2
Comparison statistics of arch shape measurements from the four models.

Model #	Computation time (min)	Arch length (m)	Arch height (m)	Cross-correlation of arch shape
Model 1	–	4.8477	0.7057	1
Model 2	130.78	4.8493	0.7217	0.9995
Model 3	29.90	4.8480	0.7180	0.9997
Model 4	1.95 (averaged for 1 ROI) 21.45 (for 11 ROIs)	4.8478	0.7164	0.9997

21.45 min for all the 11 ROIs, which indicates a further improvement on efficiency compared with Model 2.

For bridge shape extraction, all four models were projected onto the 2D evaluation plane (XY plane in Fig. 8) to generate images. Canny edge detection and circular Hough transform were implemented to identify the arch shapes and circular centroids on the deck, respectively. The bridge arch shapes are presented in Fig. 9, and the statistics are listed in Table 2. The arch length and height measurements from Model 1 were used as the reference. The three image reconstruction models provided accurate measurement of the arch length with an error < 1.6 mm. The arch height measurements deviated by 2.8, 2.9 and 1.9 mm for Models 2–4, respectively. The similarity of arch shapes, which was quantified by linear interpolation over the arch length range, exceeded 99.97% for the three models. As shown in Fig. 9, the arch edges in Model 2 have apparent data missing around X = 0.5 m probably because of inconsistent depth estimation between the red bridge arch and the top red reaction frame in the stereo fusion step. There were 23 valid images

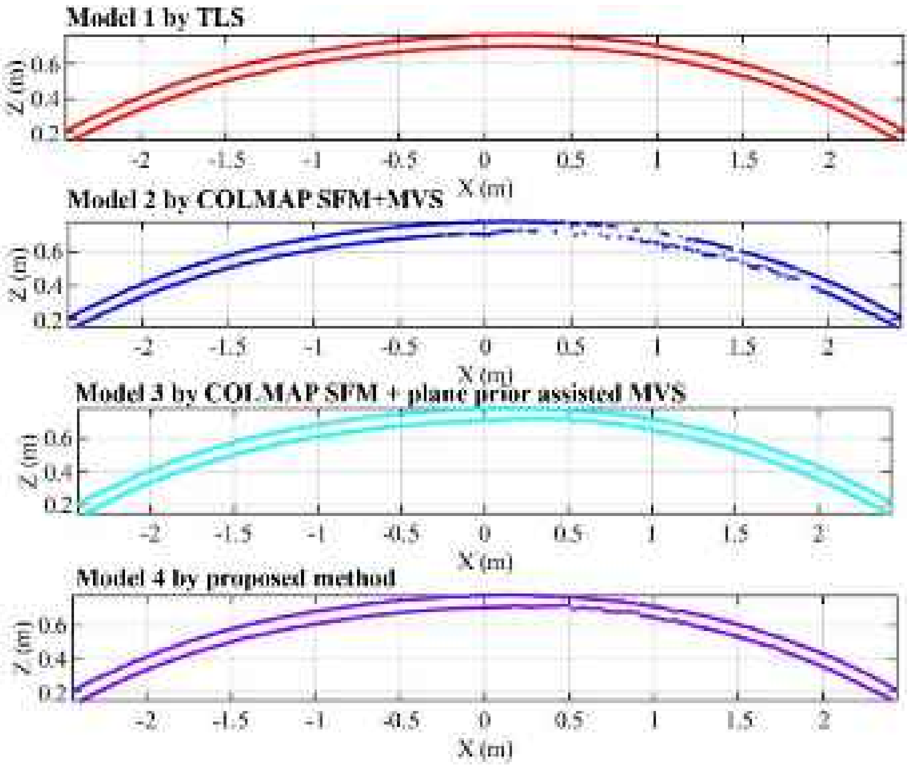


Fig. 9. Extracted arch shapes from the four point cloud models.

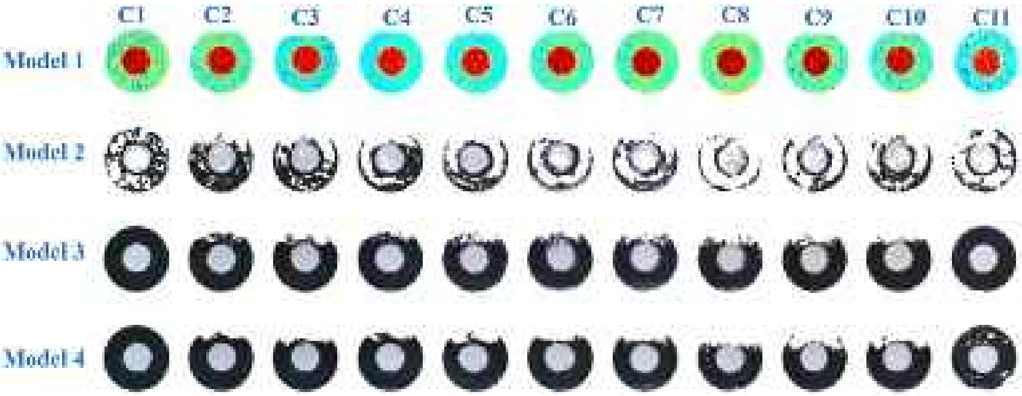


Fig. 10. Models of circular markers on bridge deck cropped from four models.

Table 3
Comparison statistics of marker measurements from the four models.

Model #	Completeness ratio (κ=1 mm)	Completeness ratio (κ=2 mm)	RMSE (mm)			
			along X	along Y	along Z	Sum
Model 1	58.9%	97.9%	0	0	0	0
Model 2	38.3%	56.6%	–	–	–	–
Model 3	81.7%	90.0%	2.272	3.198	1.790	4.312
Model 4	80.0%	86.2%	2.195	2.612	1.353	4.197

containing this region. Fifteen of these images had the bridge arch and the top reaction frame visually overlapped from their perspective views.

The planar circular markers on the bridge girder were extracted for the evaluation of model completeness. The marker regions in each model were cropped shown in Fig. 10. The ground truth model of the circular marker is a circular-shaped planar mesh with the radius of 45 mm. The model was uniformly sampled by 0.4 mm grid spacing to generate the ground-truth point cloud model for evaluation. Here, the completeness ratio [46] is defined as the percentage of points on the ground truth model that are within a specified unit spacing κ on point sets in the measured model. The unit spacing κ is taken as 1 mm and 2 mm, respectively, and the corresponding completeness ratios are presented in Table 3. Model 2 has the least completeness which are insufficient for post-processing such as extracting circular centroids for the

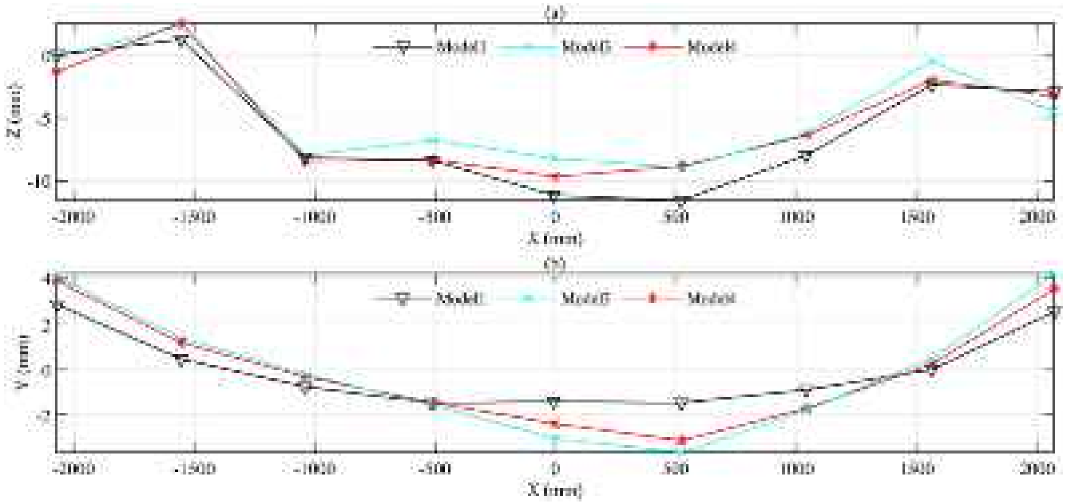
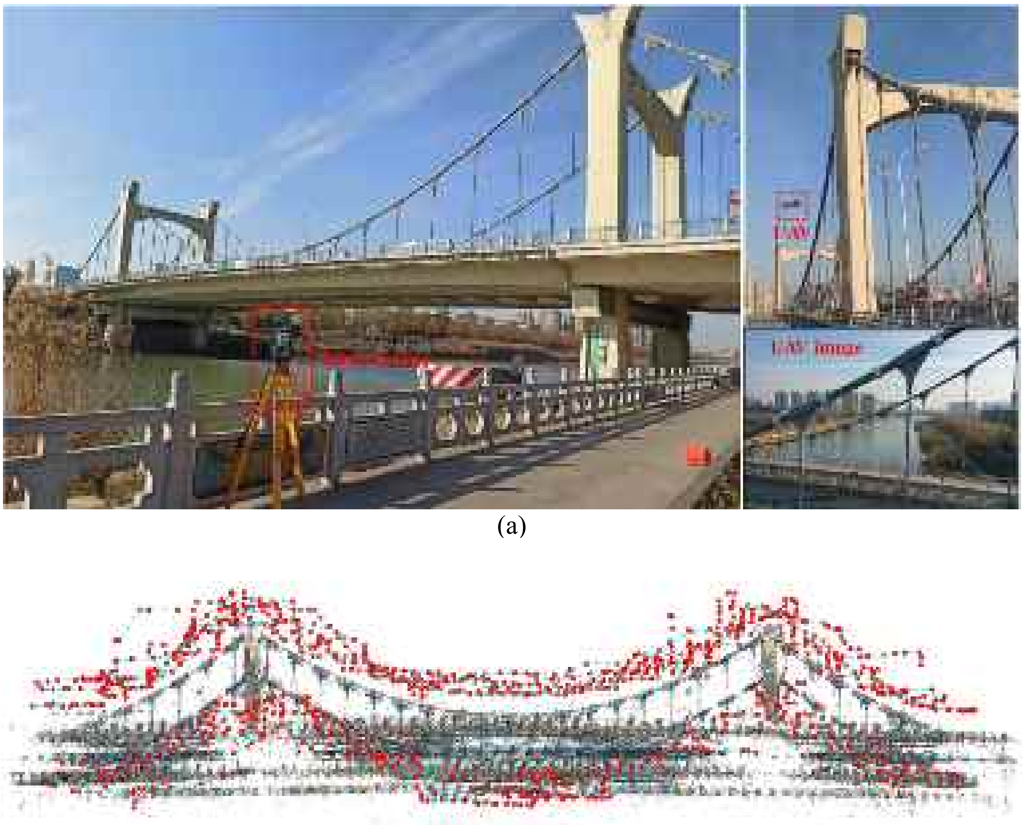


Fig. 11. Bridge deck shapes derived from marker centroids in four models: (a) along the bridge height direction (Z axis); and (b) along the out-of-plane direction (Y axis) of the girder.



(a)

(b)

Fig. 12. Figure of the tested suspension bridge and data acquisition systems (a) and the sparse point cloud model of the full bridge with the annotation of camera positions of collected UAV images (b).

markers. Model 4 achieves similar performance as Model 3, which indicates that applying sub-image clusters for distributed dense reconstruction has no apparent reduction on the model completeness. Models 3 and 4 both fails to capture the top part for ROIs 2–10 and this is due to the limitation of flying path during the test. Most of UAV image captured the bridge structure from the front top view.

The bridge girder shape is represented by the connecting curves of the detected centroids of markers C2-C10 with the measurement results presented in Fig. 11 and Table 3. The root mean square error (RMSE) along the height direction (Z axis) was on average 1.790 mm in Model 3 and 1.353 mm in Model 4. Fig. 11(b) demonstrates that the image reconstruction models (Models 3 and 4) have apparent larger out-of-plane trend for the girder compared with the scan model (Model 1) with the maximum deviation reaching 4.927 mm and 3.995 mm, respectively. This is probably because the accuracy of depth estimation for a feature is usually worse than the accuracy of localising the feature

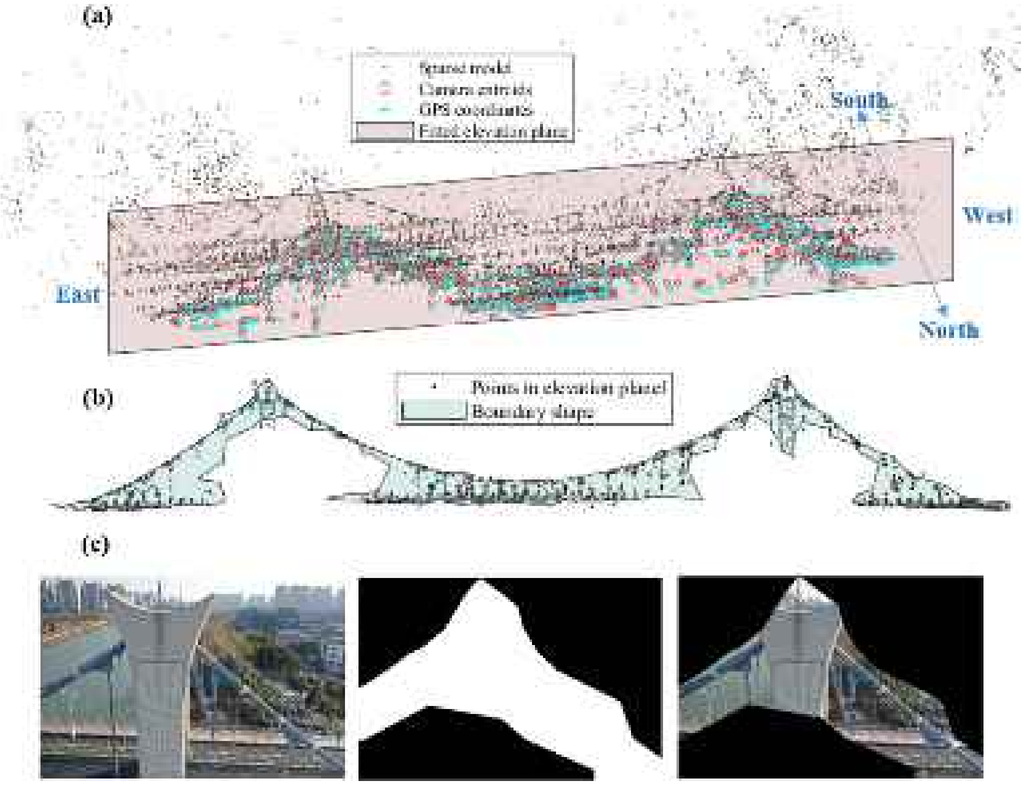


Fig. 13. Demonstration of image mask generation process for the bridge's north side: (a) sparse model aligned with the world coordinate system; (b) Boundary shape extraction from coplanar points in bridge elevation plane; and (c) an example of image mask projected from the projection of boundary shape in the image plane.

in image plane.

The laboratory validation on an arch bridge data shows the effectiveness of the proposed image reconstruction method on computational efficiency. The measurement accuracy for the deck heights is limited by approximately 1.35 mm (0.024%) with respect to the span length of 5.6 m, which is feasible for the measurement requirements on bridge application scenarios.

4. Field test on a suspension bridge for cable shape measurement

The proposed distributed image reconstruction method was implemented for suspension cable shape measurement in a suspension road bridge. The measured cable shape data were compared with those by the total station for accuracy evaluation. This validation test focused on measuring the shapes of the suspension cables because they are the main load-carrying members in the bridge system. The deck shape measurement was not considered because the bridge deck contained no apparent height difference along the span direction.

4.1. Test information

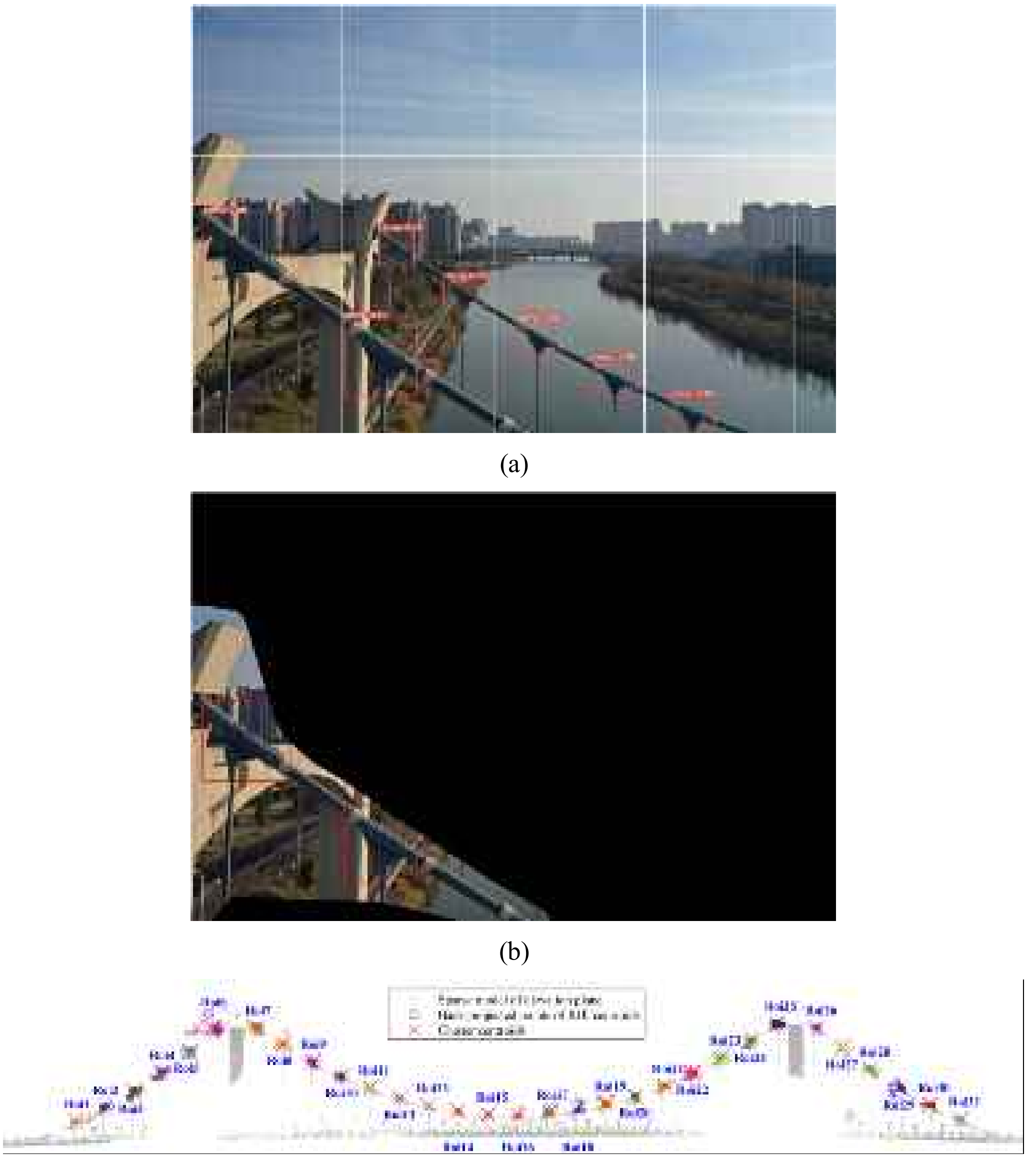
The Xiaolongwan bridge, shown in Fig. 12 (a), is a self-anchored suspension bridge located in Jiangning District, Nanjing, China. Its span length is 184 m, including a main span of 96 m and two side spans of 44 m. The height of the bridge towers is 35.4 m. The suspension cables in the main span are designed in the shape of second-order quadratic parabola with a rise-to-span ratio of 1/5.5. Because the multi-view image acquisition for full bridge coverage is not immediate, and a stable bridge status is expected during the entire acquisition process, the time range was selected as between 14.00 and 15.30 on a sunny day (22th December 2021) with relatively low-density traffic.

DJI Mavic 2 Pro was used to collect multi-view images with an image resolution of 5472 × 3648 pixels. The intrinsic parameters of the camera were pre-calibrated by observing the chessboard patterns. The flying paths are shown in Fig. 12 (b) with respect to a sparse point cloud model of the bridge and the red dots in the figure represent the camera positions used to capture images. Because the focus was on measuring the suspension cable shape, the flying paths were two independent paths for each cable plane with the UAV camera facing the bridge elevation direction. In each flying path, six routes along the bridge length direction with slightly different flying heights and distances, were arranged to ensure complete coverage. Because the suspension cables are 8.5 m inside the bridge girder edge, the flying routes were close to but outside the bridge girder edge for safety considerations. The average distance to the cable plane was calculated as 10.5 m. Given the camera focal length of 10.33 mm and the sensor resolution of 0.00234 mm/pixel, the average ground sampling distance [22] was estimated as 2.38 mm/pixel in the case with no tilt from the camera optical axis to the bridge elevation surface. The image acquisition mode was set as continuous capture at a time interval of 2 s and 1957 images were collected over a period of 80 min. The GPS coordinates extracted from the image EXIF data were analysed, and a distance threshold of 0.5 m was set to remove images with high similarity. The remaining 644 and 769 images for the north and south cable planes, respectively, were imported for 3D reconstruction analysis.

As a reference sensor, a reflectorless total station, Sokkia CX-52, was used to measure the cable shape at the same time interval. For convenience, the measured positions were set at the centroids of the pin rods under the cable clamps.

4.2. Analysis results

Because images taken from the two sides of the bridge capture non-overlapping regions of the bridge, the reconstruction process was conducted separately for the two sides. The image mask generation process for the north side of the bridge is shown in Fig. 13. An initial SFM was conducted to derive a sparse model and camera poses; these were then



(a)

(b)

(c)

Fig. 14. ROI image clustering process: (a) ROI detection in 2D image domain for an tiled image using YOLOv5; (b) masked image with the annotation of valid ROIs; and (c) clustering results of back-projected points of valid ROI centroids in 3D spatial domain.

aligned with the GPS coordinates of the images, as shown in Fig. 13(a). The bridge elevation plane was identified by RANSAC plane fitting on selected points that were < 20 m from the nearest camera positions. In-plane points were pre-processed using a statistical outlier removal filter [47] and then converted into the boundary shape based on the alpha-shape as shown in Fig. 13(b). For each image, the extracted irregular boundary shape was projected onto the image plane to generate an image mask for valid bridge regions shown in Fig. 13(c). The masked images were imported for a second-round hierarchical SFM to derive an accurate estimation of camera poses.

The ROIs containing cable clamps, were firstly detected from multi-view images using YOLOv5, as shown in Fig. 14(a). Because the images might capture the information of the two sides, the image masks were used to keep valid ROIs within the masked regions, shown in Fig. 14(c). The valid ROI centroids were back-projected onto one the bridge elevation plane to evaluate the spatial distribution by the hierarchical clustering. The threshold for the minimum Euclidean distance is set as 1.0 m (20% of the horizontal spacing for the vertical suspenders). The cluster groups with the image number below 10 were removed as outliers. The cluster results shown in Fig. 14(c) indicates 31 clusters that corresponds to 31 spatially different ROIs of cable clamps along one bridge elevation plane.

The valid regions were segmented into cropped ROI images with the annotation of cluster indices to generate 31 cropped image subsets. The average number of images in each subset is 54.3. Fig. 15 presents the image subsets for Roi8, with a total image number of 58. The camera positions from the cluster centroid varied from 8.27 to 34.54 m, and the ROI image sizes were in the range of 930×871 to 4284×3595 pixels, shown in Fig. 15(b). Instead of using all cropped images for dense reconstruction, the cropped images were sorted by the camera distance to the bridge elevation plane in ascending order, and those with the distance < 12 m were retained for reconstruction to further improve computational efficiency. If the remaining image number was <20, the closest 20 images were retained to ensure complete coverage of the ROIs. For Roi8, 26 cropped images were imported for plane-prior assisted depth map estimation and fusion, with results shown in Fig. 15(c). An enlarged-view of the cable clamp that shows the apparent colour and depth changes in the cable bands and pin rod is also included.

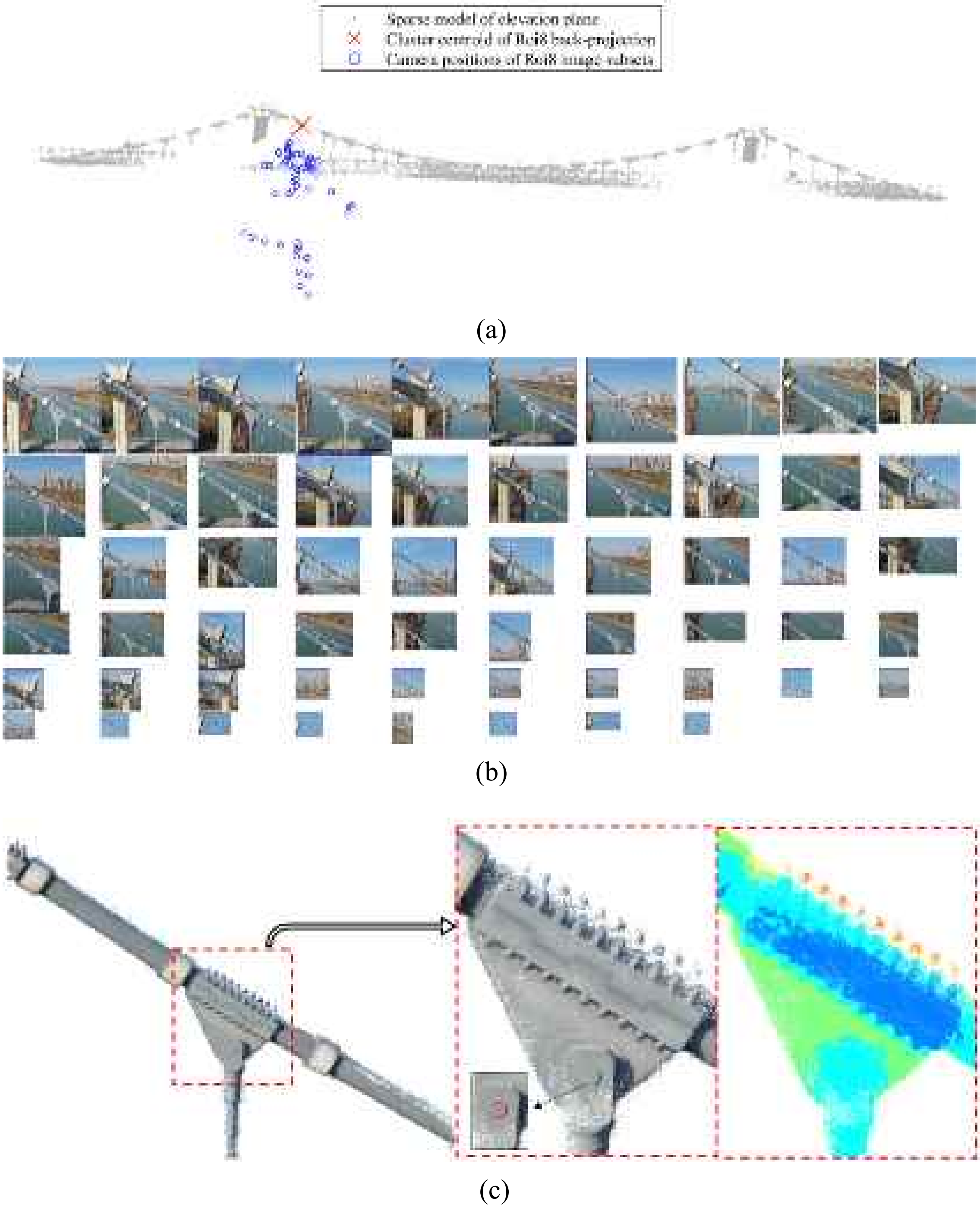

Sparse model of elevation plane
Cluster centroid of Roi8 back-projection
Camera positions of Roi8 image subsets
(a)
(b)
(c)

Fig. 15. Camera positions and cropped images for Roi8: (a) camera positions for Roi8 image subsets with respect to the bridge elevation plane; (b) cropped image subsets for Roi8; and (c) dense point cloud model for Roi8 and zoom-in view of the cable clamp region coloured by the original pattern and depth.

Fig. 16 (a) shows the image number and computation time for the dense reconstruction of each ROI. The computational efficiency depends on the number of images and image size. The average computation time for one ROI was 8.04 min in this case. Because the dense point cloud models derived for the 31 ROIs were in the same coordinate system as the previous sparse model, the 31 point cloud models were unified to a single model, and truncated to keep only the elevation side close to the camera positions, with the results shown in Fig. 16 (b). The scale factor for the model was estimated using the distance between the tops of the two towers.

To evaluate the measurement accuracy, the main cable shapes were measured by the total station on the same time for comparison. For easy consistency among multiple cable clamps, the targets for the total station were set at the small central hole edge on the pin rod for each cable clamp. The 3D coordinates on the pin rod centroids (shown in Fig. 15 (c)), were extracted from the reconstructed dense model to derive the relative heights for the main suspension cables, as shown in Fig. 17 (a). The height difference of the tower tops relative to the designed value was calculated, and then applied on the measured coordinates to recover the actual height. Fig. 17 (b) presents the measurement error of cable heights from image-based reconstruction compared with the measurements of the total station. The root mean square (RMS) errors for the

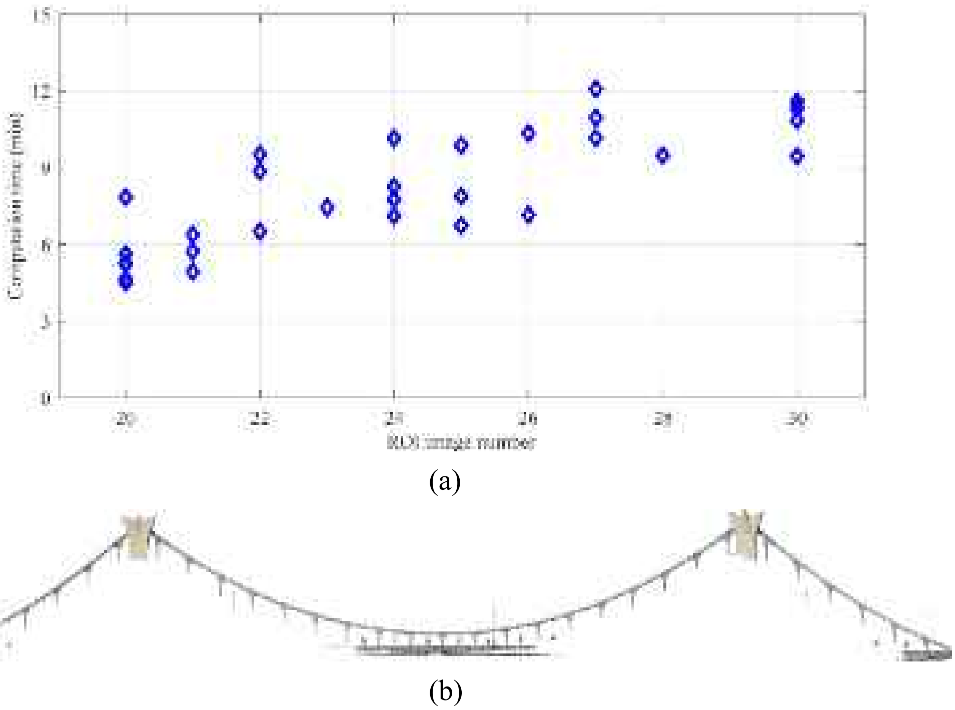


Fig. 16. Dense reconstruction results for the north side of the bridge: (a) image number and computation time for each ROI image subset; and (b) dense point cloud model for the whole bridge.

north and south sides were 1.27 cm and 1.91 cm, respectively. The maximum errors reached 2.48 cm and 4.00 cm, respectively, occurring at the end cables. Therefore, the maximum measurement error for cable heights was 0.042% relative to the main span length (96 m). The measured shapes by the image-based method were compared with the design values. The RMS deviations of the cable centroids along the horizontal and vertical directions were 1.99 cm and 5.00 cm, respectively. The height deviation proportions from the design value were < 0.54%. No severe cable shape change occurs after nine years of operation.

The requirements for the vertical positioning accuracy of the main cables are as follows: For the bridge completed status, the height of the mid-span suspension cable in the main span is allowed to deviate by < 1/20000 of the span length (equal to 4.8 mm in this case) compared with the design value [48]. For the in-service status, periodic inspection is conducted every two years including the evaluation of the suspension cable shapes that are required to fit the designed shape [49]. Consequently, image-based measurement results with centimetre-level accuracy are still unqualified for these two cases. Monthly regular inspections focus on determining the overall operation status of the bridge, mainly through visual approaches. Because bridge deterioration usually accumulates over the years, the image-based shape measurement method could be an option to supplement the existing visual inspection method by avoiding the occurrence of severe cable shape changes between periodic inspection gaps.

5. Conclusions

This paper describes a new image-based 3D reconstruction pipeline for accurate and efficient geometry measurement of bridge structures. The novelty lies in the decomposition of a full-bridge 3D reconstruction task large-scale full-size images into multiple distributed tasks of reconstructing bridge sub-models using sub-image sets. A deep learning-based object detection algorithm was implemented to automatically identify key bridge components, and then to cluster the images into cropped image subsets for each bridge component. The dense reconstruction algorithm, plane prior assisted MVS was implemented to compute the bridge sub-models using each image cluster separately; these were then merged into a full model. In addition, to improve model accuracy, an image mask generation technique based on 3D-to-2D mesh projection was used to keep valid pixels for feature correspondence search and camera pose estimation.

The proposed method was first validated in a laboratory arch bridge for arch and deck shape measurements. Incorporating plane prior assisted MVS for dense reconstruction significantly improves computational efficiency and model completeness. The proposed distributed task scheme, improves the efficiency of the dense reconstruction, requiring an average of 1.95 min for a single ROI. The derived arch shape achieved a high similarity to the TLS-derived model, with the cross-correlation coefficient exceeding 99.97%. The measurement deviations for the arch length and height were 1.6 and 1.9 mm. The measurement accuracy for the deck heights is limited by approximately 1.35 mm with respect to a span length of 5.6 m.

The method was also implemented for suspension cable shape measurements in a suspension bridge. The reconstruction of the full cable shape was partitioned into reconstructing 31 ROIs centred at the cable clamps along the bridge length direction. For a single ROI, 20 to 30 cropped images were automatically selected by YOLOv5 and the dense reconstruction process took an average of 8.04 min. The RMSE for cable height measurements was 1.91 cm compared with the total station measurements.

Although this study presented promising results, there are some limitations that require further investigation. First, the image-based reconstruction method is highly dependent on image quality, such as the image overlapping ratio and camera-to-target distance, etc. Future research should focus on developing an evaluation metric about balancing the high image quality and image acquisition efficiency. In addition, the post-processing step for quantifying bridge shapes and geometric dimensions from image-derived point cloud models lacks automation that requires further investigation. For construction quality checks and periodic inspection, the monitoring accuracy of suspension cable shapes is expected to be at the millimetre level; however, there is

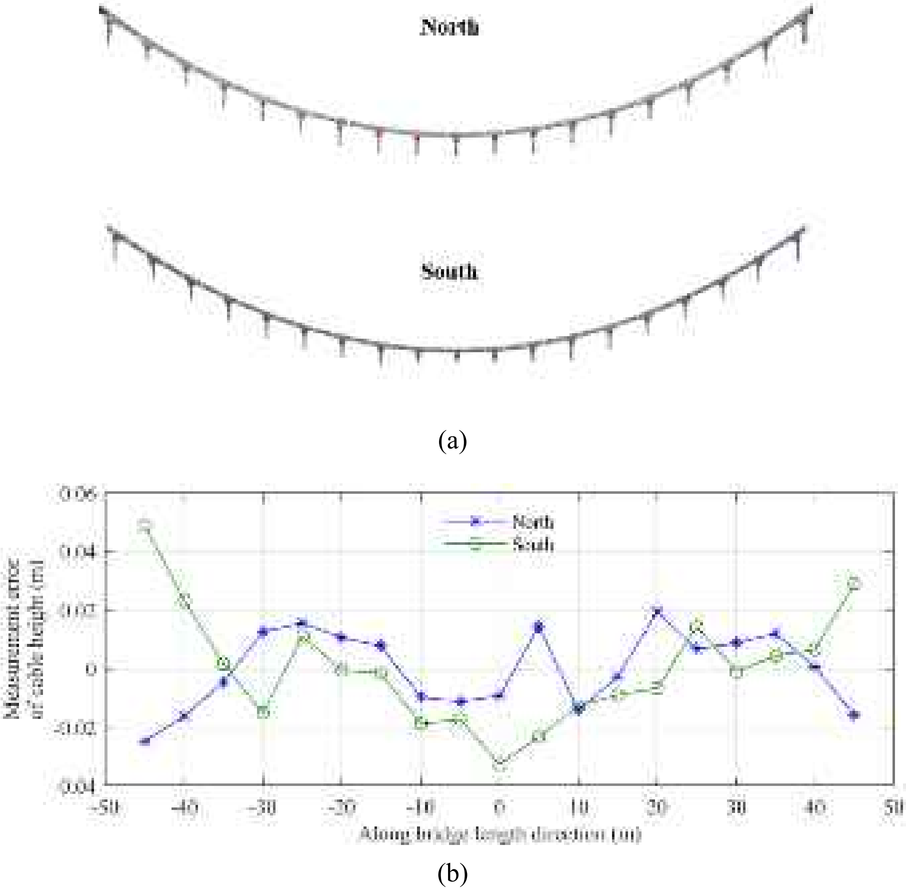


Fig. 17. Main suspension cable shape results for the north and south sides of the bridge by image-based reconstruction: (a) dense point cloud models of the main suspension cables with denotations on the pin rods; and (b) measurement error of the main cable heights from images compared with the reference height by the total station.

still a technical gap in image-based measurement accuracy in the field.

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.

Acknowledgement

This research was supported by the National Key R&D Program of China (2019YFC1511100) and the Jiangsu Natural Science Foundation (BK20190372). The authors express their sincere appreciation for their support. Also, thanks to Yi Luo and Shang Jiang for support in the laboratory and field testing; thanks to Jiangsu Kedi Construction Engineering Quality Inspection Company for support in the total station measurement on the Xiaolongwan Bridge. Finally, the authors would like to thank the two anonymous reviewers for their constructive comments.

References

[1] B.F. Spencer, V. Hoskere, Y. Narazaki, Advances in computer vision-based civil infrastructure inspection and monitoring, Engineering. 5 (2019) 199–222, https://doi.org/10.1016/j.eng.2018.11.030.

[2] Z. Shang, Z. Shen, Multi-point vibration measurement and mode magnification of civil structures using video-based motion processing, Autom. Constr. 93 (2018) 231–240, https://doi.org/10.1016/j.autcon.2018.05.025.

[3] D. Feng, T. Scarangello, M.Q. Feng, Q. Ye, Cable tension force estimate using novel noncontact vision-based sensor, Measurement. 99 (2017) 44–52, https://doi.org/10.1016/j.measurement.2016.12.020.

[4] C.Z. Dong, S. Bas, F.N. Catbas, A portable monitoring approach using cameras and computer vision for bridge load rating in smart cities, J. Civil Struct. Health Monit. 10 (2020) 1001–1021, https://doi.org/10.1007/s13349-020-00431-2.

[5] D. González-Aguilera, J. Gómez-Lahoz, Dimensional analysis of bridges from a single image, J. Comput. Civ. Eng. 23 (2009) 319–329, https://doi.org/10.1061/(ASCE)0887-3801(2009)23:6(319).

[6] B. Riveiro, D.V. Jauregui, P. Arias, J. Armesto, R. Jiang, An innovative method for remote measurement of minimum vertical underclearance in routine bridge inspection, Autom. Constr. 25 (2012) 34–40, https://doi.org/10.1016/j.autcon.2012.04.008.

[7] Q. Lu, S. Lee, Image-based technologies for constructing as-is building information models for existing buildings, J. Comput. Civ. Eng. 31 (2017) 04017005, https://doi.org/10.1061/(ASCE)CP.1943-5487.0000652.

[8] Y. Jiang, Y. Bai, Low–high orthoimage pairs-based 3D reconstruction for elevation determination using drone, J. Constr. Eng. Manag. 147 (2021) 04021097, https://doi.org/10.1061/(ASCE)CO.1943-7862.0002067.

[9] C. Popescu, B. Täljsten, T. Blanksvärd, L. Elfgren, 3D reconstruction of existing concrete bridges using optical methods, Struct. Infrastruct. Eng. 15 (2019) 912–924, https://doi.org/10.1080/15732479.2019.1594315.

[10] C. Wu, VisualSFM: A Visual Structure from Motion System. http://ccwu.me/vsfm/index.html, 2016 (accessed May 1, 2021).

[11] J.L. Schonberger, J.M. Frahm, Structure-from-motion revisited, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Las Vegas, NV, USA, 2016, pp. 4104–4113, https://doi.org/10.1109/CVPR.2016.445.

[12] D. Cernea, Multi-View Stereo Reconstruction Library. https://github.com/cdcseacave/openMVS, 2020 (accessed May 10, 2021).

[13] Agisoft. https://www.agisoft.com/, 2021 (accessed October 7, 2021).

[14] DJI, Dji Terra. https://www.dji.com/br/dji-terra, 2019 (accessed October 7, 2021).

[15] S. Jiang, C. Jiang, W. Jiang, Efficient structure from motion for large-scale UAV images: a review and a comparison of SfM tools, ISPRS J. Photogramm. Remote Sens. 167 (2020) 230–251, https://doi.org/10.1016/j.isprsjprs.2020.04.016.

[16] M. Pollefeys, Visual 3D Modeling from Images, Chapel Hill, USA, 2004. https://www.cvg.ethz.ch/teaching/compvis/2012/tutorial.pdf.

[17] H. Fathi, F. Dai, M. Lourakis, Automated as-built 3D reconstruction of civil infrastructure using computer vision: achievements, opportunities, and challenges, Adv. Eng. Inform. 29 (2015) 149–161, https://doi.org/10.1016/j.aei.2015.01.012.

[18] M. Golparvar-Fard, F. Peña-Mora, S. Savarese, Integrated sequential as-built and as-planned representation with D4AR tools in support of decision-making tasks in the AEC/FM industry, J. Constr. Eng. Manag. 137 (2011) 1099–1116, https://doi.org/10.1061/(ASCE)CO.1943-7862.0000371.

[19] N. Kassotakis, V. Sarhosis, M.V. Peppa, J. Mills, Quantifying the effect of geometric uncertainty on the structural behaviour of arches developed from direct measurement and Structure-from-Motion (SfM) photogrammetry, Eng. Struct. 230 (2021), 111710, https://doi.org/10.1016/j.engstruct.2020.111710.

[20] D. Moon, S. Chung, S. Kwon, J. Seo, J. Shin, Comparison and utilization of point cloud generated from photogrammetry and laser scanning: 3D world model for smart heavy equipment planning, Autom. Constr. 98 (2019) 322–331, https://doi.org/10.1016/j.autcon.2018.07.020.

[21] G. Morgenthal, N. Hallermann, J. Kersten, J. Taraben, P. Debus, M. Helmrich, V. Rodehorst, Framework for automated UAS-based structural condition assessment of bridges, Autom. Constr. 97 (2019) 77–95, https://doi.org/10.1016/j.autcon.2018.10.006.

[22] S. Chen, D.F. Laefer, E. Mangina, S.M.I. Zolanvari, J. Byrne, UAV bridge inspection through evaluated 3D reconstructions, J. Bridg. Eng. 24 (2019) 05019001, https://doi.org/10.1061/(ASCE)BE.1943-5592.0001343.

[23] I.G. Dino, A.E. Sari, O.K. Iseri, S. Akin, E. Kalfaoglu, B. Erdogan, S. Kalkan, A. A. Alatan, Image-based construction of building energy models using computer vision, Autom. Constr. 116 (2020), 103231, https://doi.org/10.1016/j.autcon.2020.103231.

[24] K. Bacharidis, F. Sarri, V. Paravolidakis, L. Ragia, M. Zervakis, Fusing georeferenced and stereoscopic image data for 3D building Façade reconstruction, ISPRS Int. J. Geo Inf. 7 (2018) 151, https://doi.org/10.3390/ijgi7040151.

[25] Y. Pan, Y. Dong, D. Wang, A. Chen, Z. Ye, Three-dimensional reconstruction of structural surface model of heritage bridges using UAV-based photogrammetric point clouds, Remote Sens. 11 (2019) 1204, https://doi.org/10.3390/rs11101204.

[26] F. Hu, J. Zhao, Y. Huang, H. Li, Structure-aware 3D reconstruction for cable-stayed bridges: a learning-based method, computer-aided civil and infrastructure, Engineering. 36 (2021) 89–108, https://doi.org/10.1111/mice.12568.

[27] Y. Xu, X. Shen, S. Lim, X. Li, Three-dimensional object detection with deep neural networks for automatic as-built reconstruction, J. Constr. Eng. Manag. 147 (2021) 04021098, https://doi.org/10.1061/(ASCE)CO.1943-7862.0002003.

[28] A. Rashidi, E. Karan, Video to BrIM: automated 3D as-built documentation of bridges, J. Perform. Constr. Facil. 32 (2018) 04018026, https://doi.org/10.1061/(ASCE)CF.1943-5509.0001163.

[29] A. Khaloo, D. Lattanzi, Hierarchical dense structure-from-motion reconstructions for infrastructure condition assessment, J. Comput. Civ. Eng. 31 (2017) 04016047, https://doi.org/10.1061/(asce)cp.1943-5487.0000616.

[30] A. Khaloo, D. Lattanzi, A. Jachimowicz, C. Devaney, Utilizing UAV and 3D computer vision for visual inspection of a large gravity dam, Front. Built Environment. 4 (2018) 1–16, https://doi.org/10.3389/fbuil.2018.00031.

[31] B. Xu, C. Liu, A 3D reconstruction method for buildings based on monocular vision, Computer-Aided Civil Infrastruct. Eng. 37 (2022) 354–369, https://doi.org/10.1111/mice.12715.

[32] J.L. Carrivick, M.W. Smith, D.J. Quincey, Structure from Motion in the Geosciences, John Wiley & Sons, Ltd, Chichester, UK, 2016, https://doi.org/10.1002/9781118895818.

[33] N. Saovana, N. Yabuki, T. Fukuda, Development of an unwanted-feature removal system for structure from motion of repetitive infrastructure piers using deep learning, Adv. Eng. Inform. 46 (2020), 101169, https://doi.org/10.1016/j.aei.2020.101169.

[34] N. Saovana, N. Yabuki, T. Fukuda, Automated point cloud classification using an image-based instance segmentation for structure from motion, Autom. Constr. 129 (2021), 103804, https://doi.org/10.1016/j.autcon.2021.103804.

[35] Y. Narazaki, V. Hoskere, T.A. Hoang, Y. Fujino, A. Sakurai, B.F. Spencer, Vision-based automated bridge component recognition with high-level scene consistency, Computer-Aided Civil Infrastruct. Eng. 35 (2020) 465–482, https://doi.org/10.1111/mice.12505.

[36] A. Mirzazade, C. Popescu, T. Blanksvärd, B. Täljsten, Workflow for off-site bridge inspection using automatic damage detection-case study of the Pahtajokk bridge, Remote Sens. 13 (2021) 2665, https://doi.org/10.3390/rs13142665.

[37] J. Dong, S. Soatto, Domain-size pooling in local descriptors: DSP-SIFT, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Boston, MA, USA, 2015, pp. 5097–5106, https://doi.org/10.1109/CVPR.2015.7299145.

[38] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography, Readings Comput. Vision. (1987) 726–740, https://doi.org/10.1016/b978-0-08-051581-6.50070-2.

[39] V. Lepetit, F. Moreno-Noguer, P. Fua, EPnP: an accurate O(n) solution to the PnP problem, Int. J. Comput. Vis. 81 (2009) 155–166, https://doi.org/10.1007/s11263-008-0152-6.

[40] Q. Xu, W. Tao, Planar prior assisted patchmatch multi-view stereo, in: Proceedings of the AAAI Conference on Artificial Intelligence, New York, USA, 2020, pp. 12516–12523, https://doi.org/10.1609/aaai.v34i07.6940.

[41] G. Jocher, A. Stoken, J. Borove, A. Chaurasia, T. Xie, L. Changyu, ultralytics/yolov5: v5.0, 2021, https://doi.org/10.5281/zenodo.4679653.

[42] J.L. Schönberger, E. Zheng, J.-M. Frahm, M. Pollefeys, Pixelwise view selection for unstructured multi-view stereo, in: European Conference on Computer Vision, Amsterdam, The Netherlands, 2016, pp. 501–518, https://doi.org/10.1007/978-3-319-46487-9_31.

[43] H. Edelsbrunner, D. Kirkpatrick, R. Seidel, On the shape of a set of points in the plane, IEEE Trans. Inf. Theory 29 (1983) 551–559, https://doi.org/10.1109/TIT.1983.1056714.

[44] J. Redmon, A. Farhadi, YOLOv3: An Incremental Improvement. http://arxiv.org/abs/1804.02767, 2018.

[45] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, YOLOv4: Optimal Speed and Accuracy of Object Detection. http://arxiv.org/abs/2004.10934, 2020.

[46] S.M. Seitz, B. Curless, J. Diebel, D. Scharstein, R. Szeliski, A comparison and evaluation of multi-view stereo reconstruction algorithms, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE, New York, NY, USA, 2006, pp. 519–526, https://doi.org/10.1109/CVPR.2006.19.

[47] R.B. Rusu, Z.C. Marton, N. Blodow, M. Dolha, M. Beetz, Towards 3D point cloud based object maps for household environments, Robot. Auton. Syst. 56 (2008) 927–941, https://doi.org/10.1016/j.robot.2008.08.005.

[48] CJJ, Code for Construction and Quality Acceptance of Bridge Works in City, Industrial Standard of the People's Republic of China, CJJ 2–2008, China Construction Industry Press, Beijing China, 2018.

[49] JTG, Technical Specifications for Maintenance of Highway Cable-Supported Bridge, Industrial Standard of the People's Republic of China, JTG/T 5122–2021, People's Transportation Press, Beijing China, 2021.