Smart Structures and Systems, *Vol. 30, No. 5 (2022) 501-511*
https://doi.org/10.12989/sss.2022.30.5.501

Deep learning approach to generate 3D civil infrastructure models using drone images

Ji-Hye Kwon [1a], Shekhroz Khudoyarov [2b], Namgyu Kim [3c] and Jun-Haeng Heo[*1]

[1] *School of Civil and Environmental Engineering, Yonsei University, 50 Yonsei-ro Seodaemun-gu, Seoul, 03722, Republic of Korea*
[2] *SISTech Co., LTD, 209, Neungdong-ro, Gwangjin-gu, Seoul, 05006, Republic of Korea*
[3] *Research Strategic Planning Department, Korea Institute of Civil Engineering and Building Technology, 283, Goyangdae-Ro, Ilsanseo-Gu, Goyang-Si, Gyeonggi-Do, 10223, Republic of Korea*

(*Received July 20, 2022, Revised September 1, 2022, Accepted September 12, 2022*)

Abstract. Three-dimensional (3D) models have become crucial for improving civil infrastructure analysis, and they can be used for various purposes such as damage detection, risk estimation, resolving potential safety issues, alarm detection, and structural health monitoring. 3D point cloud data is used not only to make visual models but also to analyze the states of structures and to monitor them using semantic data. This study proposes automating the generation of high-quality 3D point cloud data and removing noise using deep learning algorithms. In this study, large-format aerial images of civilian infrastructure, such as cut slopes and dams, which were captured by drones, were used to develop a workflow for automatically generating a 3D point cloud model. Through image cropping, downscaling/upscaling, semantic segmentation, generation of segmentation masks, and implementation of region extraction algorithms, the generation of the point cloud was automated. Compared with the method wherein the point cloud model is generated from raw images, our method could effectively improve the quality of the model, remove noise, and reduce the processing time. The results showed that the size of the 3D point cloud model created using the proposed method was significantly reduced; the number of points was reduced by 20-50%, and distant points were recognized as noise. This method can be applied to the automatic generation of high-quality 3D point cloud models of civil infrastructures using aerial imagery.

Keywords: automatic model generation; deep learning algorithm; noise reduction; point cloud; semantic segmentation

1. Introduction

In recent years, owing to the advancement in drones, sensors, and software, the generation of three-dimensional (3D) models for civil infrastructures has become more affordable, increasing its popularity. Light detection and ranging (LiDAR) sensors are commonly used to obtain point cloud data from civil infrastructures. However, point cloud data scanned by LiDAR are very heavy and noisy. Thus, more time and effort is required to extract the region of interest (ROI) by manually removing the noise. Alternatively, the generation of 3D models for civil infrastructures can be performed using photogrammetry techniques based on 2D images. 2D images are widely used in the visual inspection of civil infrastructures; thus, they are preferable to LiDAR data. Moreover, with the help of image-based artificial intelligence techniques, the automation of the visual inspection of civil infrastructures can be rapidly accelerated.

With the advent of unmanned aerial vehicles (UAV) or drones, 3D reconstruction using photogrammetry has become a valuable tool for numerous civil engineering applications, including the preservation of historical buildings (Khaloo *et al*. 2018); structural modeling (Popescu *et al*. 2019, Tang *et al*. 2019); road construction (Inzerillo *et al*. 2018); structural assessment and progress measurement (El-Omari and Moselhi 2008); investigation of land usage (Liu and Abd-Elrahman 2018, Jiang *et al*. 2020); and monitoring of construction site activities (Omar *et al*. 2018). Irschara *et al*. (2010) discussed the automation of photogrammetry reconstruction using digital images captured from UAVs. They demonstrated the feasibility of accurate and quick reconstructing 3D scenes of buildings from unordered images captured using a UAV platform. Their system comprises three processing steps: feature extraction, matching, and structure from motion (SfM) computation. He *et al*. (2022) quantified the spatial distribution of the inter-rill and rill erosion of losses on different slopes using SfM photogrammetry. They used SfM photogrammetry to measure slope surfaces to collect data in a slope erosion simulation. The photogrammetry approach provided a 3D point cloud as a complete and quantitative documentation of civil objects at a given moment without any subjective evaluations by field investigators. The 3D point cloud provides additional geometric and geological information for analyzing the stability of the structure (Menegoni *et al*. 2019). The 3D reconstruction was also helpful for drones in

*Corresponding author, Ph.D., Professor,
E-mail: jhheo@yonsei.ac.kr
[a] Ph.D. Candidate, E-mail: wisegruv@gmail.com
[b] Ph.D., E-mail: shekhrozx@gmail.com
[c] Ph.D., E-mail: namgyu.kim@kict.re.kr

ISSN: 1738-1584 (Print), 1738-1991 (Online)

dam applications as the drones could survey a wide range of steep dams effectively without including the risk of climbing faced by field investigators. A similar method was used to obtain point cloud spatial information in various applications, such as the 3D documentation of cultural heritage sites (Rahaman and Champion 2019). Unsupervised reconstruction of complete temporally coherent 4D scene models with improved non-rigid objects segmentation and shape reconstruction method is proposed. The work contributes an automatic method for initial coarse reconstruction to initialize joint estimation; sparse-to-dense temporal correspondence integrated with joint multi-view segmentation and reconstruction of dynamic scenes by introducing shape constraint (Mustafa *et al*. 2021). Segmentation and geometric features based on deep learning, as well as the creation of a semantic 3D reconstruction using map stitching using captured images by the eye-in-hand vision system, were proposed by other researchers (Zha *et al*. 2020). The work shows that the quality of segmented images and the accuracy of 3D semantic reconstruction are effectively improved. Other researchers have proposed an automatic method that generates accurate segmentation of 2D and 3D objects from light fields. Their method makes good use of high-density data coherence and works efficiently with thousands of input images (Yücer *et al*. 2016). Liu *et al*. (2022) proposed a segmentation method combined with photogrammetry to build a 3D segmented model to measure biological growth area. The main findings of this work are that image segmentation can be easily applied to analyze crowdsourced data, and photogrammetry can be combined with image segmentation to analyze images for monitoring purposes.

The noise in 3D models generated from raw drone images is unavoidable because these images include unwanted areas such as vegetation, trees, and the sky. This study presents a procedure for the automatic generation of 3D models for civil infrastructures using deep learning and photogrammetry techniques. First, ROIs can be extracted from drone images using semantic segmentation. Minaee *et al*. (2021) presented an overview of image segmentation frameworks and models. In this study, we evaluated several such deep learning models based on their performance, accuracy, speed, and storage requirements. Following the survey, DeepLab (Chen *et al*. 2014, 2017a, b, 2018) was used in this study because it achieved the best performance in a similar dataset, Cityscapes (2020). Then, a 3D model was generated from the segmented images using a 3D reconstruction software. Although there are several commercial 3D reconstruction software packages such as Agisoft (2021) and Pix4D (2021), this study used an open-source 3D modeling pipeline, COLMAP (2022), which offered command-line options essential to building the proposed automatic modeling procedure. In addition, the COLMAP generation procedure produced nearly the same results that Pix4D did, with denser point clouds. Comparing the process of generation with the proposed algorithm showed that the volume of the point cloud by the proposed algorithm with segmented images, decreased by almost 20%, and the process took 10 min less than the reconstruction process with the original dataset.

The remainder of this paper is organized as follows. Section 2 describes the theoretical background of the 3D reconstruction using photogrammetric techniques. Section 3 describes the noise-reduction methodology used for generating a 3D model of civil infrastructures. Section 4 details the implementation of the proposed methodology using experimental field data from Korea. Section 5 concludes the study with a summary and discussion of the applications.

2. Theoretical background

2.1 Semantic segmentation

The goal of image segmentation is to simplify and change the representation of an image into something more meaningful and easier to analyze. More precisely, this involves labeling every pixel of an image and assigning it into one of the known classes. 2D images with known classes of unique colors can be used as the output of an image segmentation algorithm. There are many diverse applications of image segmentation such as the medical analysis of human organs, the autopilot function in self-driving cars, and robotic vision.

Image segmentation can be classified into two main techniques: semantic and instance. Semantic segmentation assigns each pixel of an image into a class, whereas instance segmentation requires a further detection of object instances. Fig. 1 shows the results of semantic segmentation on a slope. Since the structural health monitoring method focuses on just a single piece of civil infrastructure, semantic segmentation can be used here. Recently, deep learning-based models have been popular for semantic segmentation in computer vision applications.

Supported by the computing power and performance of deep learning, semantic segmentation has recently become available for applications in civil engineering tasks, such as crack detection and pavement monitoring. DeepLab (Chen *et al*. 2017a) is a convolutional model that introduces dilated convolution, which introduces a parameter called the dilation rate for the convolutional layers; the dilation rate defines the spacing between the kernel weights. Fig. 2 shows the dilated convolution process with several dilation rates. This approach was inspired by the success of the R-CNN spatial pyramid pooling method, which showed that regions of an arbitrary scale could be accurately and efficiently classified by resampling the convolutional features extracted at the same scale. A variant of the scheme

(a) Input image (b) Semantic segmentation (c) Extracted ROI

Fig. 1 ROI extraction results

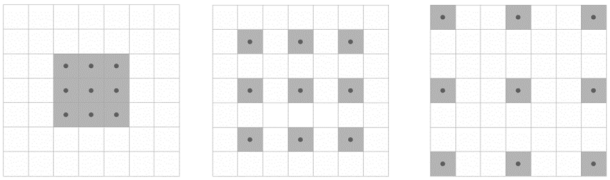


Fig. 2 Dilated convolution process with different dilation rates on 3×3 kernel

that used several parallel convolutional layers with different sampling rates was implemented. The features extracted for each sample rate were further processed into separate branches and combined to obtain the final result.

The dilated convolution output y(i) on an input of x(i) is given in Eq. (1)

$$y(i) = \sum_{k=1}^{K} x[i + rk]w[k] \tag{1}$$

where r is the dilation rate and w is the kernel weight (El-Omari and Moselhi 2008). Dilated convolution enables deep learning models to capture more spatial features without computational overhead. Conditional random fields (CRFs) have traditionally been employed to smoothen noisy segmentation maps. Typically, these models couple neighboring nodes, favoring same-label assignments to spatially proximal pixels. Qualitatively, the primary function of these short-range CRFs is to clean up the spurious predictions of weak classifiers built on top of local hand-engineered features. The model employs an energy function

$$E(x) = \sum_{i} \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \tag{2}$$

where x is the label assignment for the pixels. And $\theta_i(x_i) = -logP(x_i)$ as the unary potential, where $P(x_i)$ is the label assignment probability at pixel i as computed by a deep CNN.

DeepLab models are known to use this technique to efficiently segment images with higher mean intersection over union (mIoU) scores than those of images segmented under techniques used by other models (Minaee *et al.* 2021). There are several versions of DeepLab models: DeepLabv1, DeepLabv2 (Chen *et al.* 2017a), DeepLabv3 (Chen *et al.* 2017b), and DeepLabv3+ (Chen *et al.* 2018). According to Minaee et al. (2021), DeepLabv3+ achieved mIoU scores of 89.0 and 82.1% on the PASCAL VOC and Cityscapes datasets, respectively; thus, it was used as the image segmentation model in this study.

Importantly, this model was amenable to efficient approximate probabilistic inference. Message-passing updates under a fully decomposable mean-field approximation

$$b(x) = \prod_{i} b_i(x_i) \tag{3}$$

can be expressed as Gaussian convolutions in bilateral spaces. High-dimensional filtering algorithms significantly sped-up this computation, resulting in an algorithm that was practically very fast, requiring less than 0.5 s on average for Pascal VOC images using the publicly available implementation.

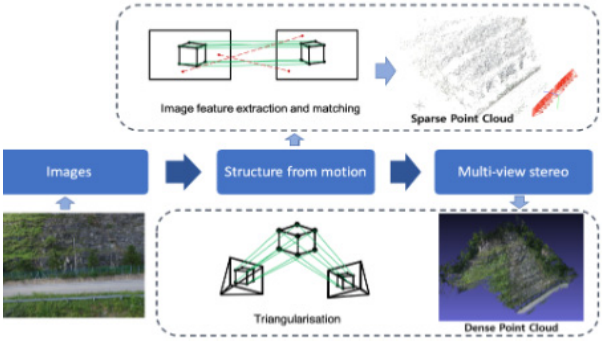


Fig. 3 Photogrammetry procedure

2.2 Photogrammetry

Photogrammetry is a technique used to obtain reliable data on real-world objects in the environment by creating 3D models from photos. 2D and 3D data are extracted from an image; these data and the overlapping images of an object, building, or terrain are then converted into a digital 3D model. This model can capture large structures and even landscapes, which would otherwise be infeasible to scan. Photogrammetry is often used by surveyors, architects, engineers, and contractors to create topographic maps, networks, and point clouds.

The photogrammetry procedure comprises two steps: SfM and multi-view stereo (MVS), as illustrated in Fig. 3. The SfM also determines the 3D coordinates of the feature points, which is known as a sparse point cloud. Since a sparse point cloud is usually not dense enough to represent an object, building, or terrain, an MVS procedure is performed to generate a dense point cloud by triangulation on each pixel of the two spatially adjacent images; this determines its corresponding 3D coordinates in the world coordinates.

Unlike traditional photogrammetry, the camera positions derived from SfM lack the scale and orientation provided by ground control points (GCPs). Consequently, the 3D point clouds are generated in a relative "image-space" coordinate system, which needs to be aligned to a real-world "object-space" coordinate system. In most cases, the transformation of SfM image-space coordinates into an absolute coordinate system can be achieved using a 3D similarity transform based on a small number of known GCPs with known object-space coordinates.

SfM estimates the camera position and orientation of each image (i.e., the structure) by triangulating the feature points appearing on multiple images. The coordinate accuracy based on the triangulation method can be compared to the accuracy of the geodetic coordinates of the bridging point network. The most analytic aerial triangulation methods, which have single and double models or photographs as basic units, are based on the vector ray coplanarity principle, given as

$$\begin{vmatrix} b_X & b_Y & b_Z \\ A'_X & A'_Y & A'_Z \\ A''_X & A''_Y & A''_Z \end{vmatrix} = 0 \quad (4)$$

where A is object coordinates of point A, b is calibrated principal point position and focal length of the camera, and the vector ray collinearity principle, given as

$$\begin{bmatrix} X - X_0 \\ Y - Y_0 \\ Z - Z_0 \end{bmatrix} = \lambda \cdot \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \cdot \begin{bmatrix} x' \\ y' \\ -f \end{bmatrix} \quad (5)$$

where X, Y, Z are ground control points, X_0, Y_0, Z_0 are object coordinates of the perspective center, λ-scale, R is three-dimensional rotation matrix, x, y, f are calibrated principal point position and principal distance of the camera.

The results derived from using the collinearity principle, wherein a photograph was the basic unit, were more rigorous, but they required larger computations.

The photogrammetric MVS workflow is as follows: a) design of the imaging configuration, which includes providing targets, imaging the network configuration, and selecting the sensor; b) geometric and radiometric calibration of the camera(s); c) capture and removal of geometric distortions from the images; d) accurate image measurements using SfM; e) correct scaling and improvement of accuracy using a photogrammetric bundle adjustment; f) image clustering and selection of images with the best content for reconstruction; g) generation of a dense 3D point cloud with MVS methods; and h) surface reconstruction and rendering. The MVS usually increases the density of the point cloud by at least two orders of magnitude.

3. Guidelines for capturing drone images

The basic components for capturing images using drones for civil infrastructures are as follows.

Ground control points (GCPs): They are known coordinates fixed in the building structure prior to the data collection process. The distances between the control points in the real world can be proportionally used to analyze the accuracy of the 3D models. GCPs help convert point-cloud data distances into real distances. Fig. 4. shows an image of the GCPs used in the laboratory experiments in this study.

Image overlapping ratio: This is calculated using the overlapping area of two consecutively captured images, as illustrated in Fig. 5(a). The translation distance between the two frames determines the overlapping ratio; thus, after image capture, the location of the next point can be determined based on the recommended overlapping ratio for the structure.

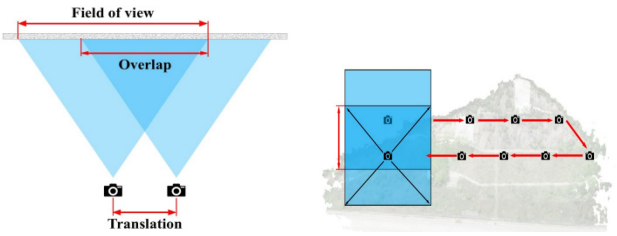
There are two types of overlapping areas in images captured by drones: ***forward*** and ***side***. The forward overlapping area is that between two consecutive frames along the flight path of the drone. Drones capture images along a linear path of the surface of the structure, and the side overlapping ratio determines the distance needed for moving to the following line along the image-capture path, as shown in Fig. 5(b).

However, commercially available photogrammetry-based 3D reconstruction software prefers other overlapping ratios. Agisoft (2021) recommended a 60% side overlap and 80% straight overlap in their documentation. Similarly, Pix4D prefers 60% side overlap and 75% front overlap for normal use (2021).

Camera angle: The angle of vision needs to be horizontal to the surface of the civil infrastructural element as drones fly along its trajectory. It is recommended that instead of rotating it, the camera should be moved to the next location to capture more scenes, as suggested by the COLMAP documentation as shown in Fig. 6.

Drone flight plan: When planning the drone's flight, the shape of the structure or objects should be considered, which includes the territory of the structure, fences, type of surface of the objects, protective nets, and roads. The drones follow the trajectory of the grid to capture images, as shown in Fig. 7.

80% forward and 60% side overlapping ratios were recommended for cut-slopes. The grid trajectory could start from either the top or bottom. For example, consider that the horizontal field view is 50 m on the surface of the cut structure. In this case, the drone should fly 10 m to the left or right to capture another image and maintain an 80%



(a) Forward overlapping ratio (b) Side overlapping ratio

Fig. 5 Overlapping ratio

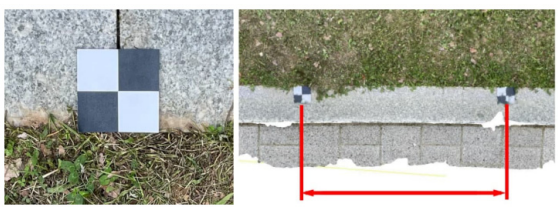


Fig. 4 GCPs in experiments

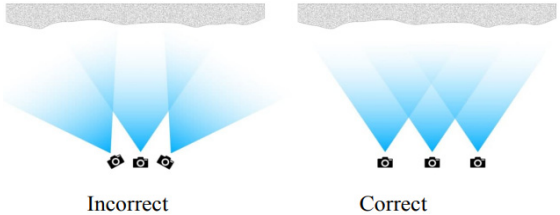


Fig. 6 Camera's angle of view

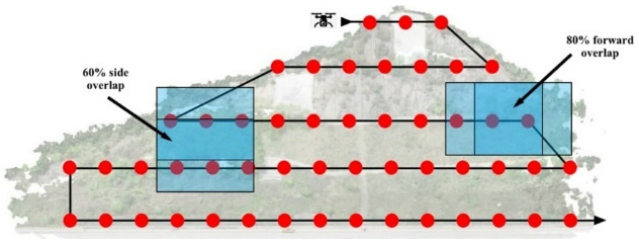


Fig. 7 Drone's flight plan with overlapping ratios

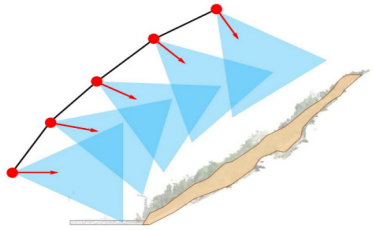


Fig. 8 Calibration of camera angle by 15° on each line

forward overlapping ratio. Similarly, the vertical field of view of the camera provides the vertical translation of the drone after each line of the flight trajectory. It is recommended to calibrate the camera angle as the drone changes trajectory line to capture the proper design of the protective structures. Fig. 8 shows the proposed plan for calibrating the camera angle for cut-slope objects.

The structures of dam components are different from those of the cut-slopes. Dam components contain straighter structures, corners, and vertical objects. Accordingly, this study proposes different drone-capture methods and flight plans. The drone needs to capture with a smaller angle of rotation on the upstream and downstream sides of the dam, corresponding to larger angles in the crest section. A horizontal surface was present on the top of the dam. Fixing a small angle of rotation is sometimes insufficient for obtaining more detailed visual information. Capturing a larger angle of rotation can provide more detail to the inner corners of the top section, as shown on the left side in Fig. 9. Experiments showed that capturing the spillway of a dam at an angle perpendicular to the ground was inefficient. The sidewalls of the spillway were located vertically. An image captured at 90° only included the horizontal surface of the spillway, excluding the sidewalls of the spillway, leading to a possible loss of data on the sidewall surface. This study proposes capturing spillways with 80% front and 60% side overlapping ratios and 30–60° on the sides of the spillway, as shown on the right side in Fig. 9.

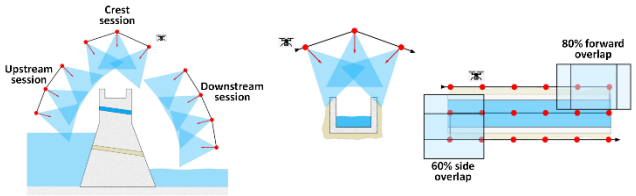


Fig. 9 Capture requirements of drone for the side of the dam. Various dam components are shown

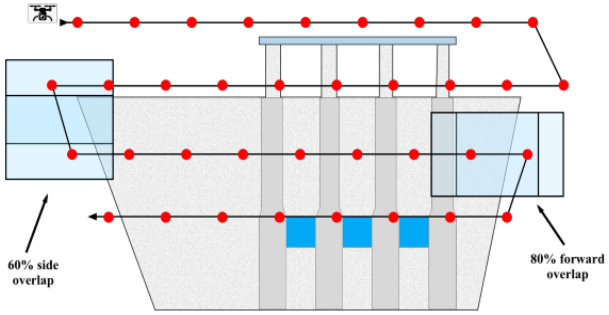


Fig. 10 Drone capture-flight plan with overlapping ratios

Following the same principle, 80% front and 60% side overlapping ratios have been recommended when building drone capture paths throughout the dam structure, as illustrated in Fig. 10.

4. 3D model generation using drone images

Currently, drones with high-resolution digital cameras, which bypass the scanning step, are available at reasonable prices. These cameras are classified into metric and amateur cameras. Close-range digital cameras are becoming increasingly popular owing to their economic advantages. The drone used in this study was a DJI T600 Inspire 1 model with a Zenmuse X3-FC350 digital camera. The camera's maximum resolution was 4864 × 3648 px, and its entire resolution was used for capturing the civil infrastructures. Before using a drone and camera, the camera needs to be calibrated to meet the required photographic conditions. Reconstructing a 3D structure from images is easier when calibrated cameras are used; in such cameras, the mapping between the image coordinates and directions relative to the camera is known. It is also necessary to plan the drone-capture trajectory and adjust the drone accordingly.

In this study, the proposed procedure for generating 3D models for civil infrastructure contains six main steps: pre-processing (downscaling images); semantic segmentation; post-processing (upscaling images); extraction of ROIs using segmentation results; generation of point cloud data; and scale adjustment; as illustrated in Fig. 11.

- **Pre-processing (downscaling images).** Experiments have shown that downsizing the size of raw drone images is important for saving computational time during the segmentation process, significantly increasing the speed of the process, and delivering better results. Median Blurring non-linear filtering technique and optimization methods with regularization conditions have been applied to preserve key points for photogrammetry. Median Blurring non-linear filter is an effective technique in reducing a certain type of noise with considerably less edge blurring as compared to the linear filters of the same size. The role of regularization conditions was to reduce the test or generalization error without affecting the initial training error.

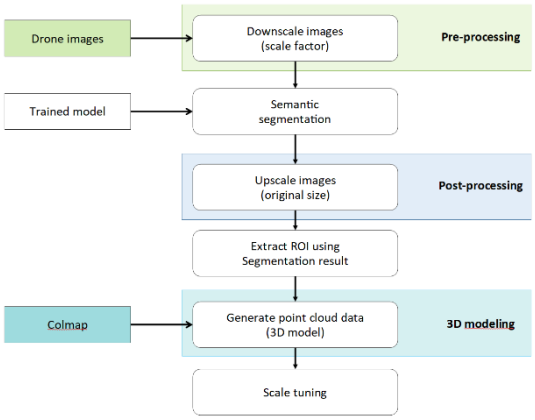


Fig. 11 Procedure for automatic generating 3D models using drone images

- **Semantic segmentation**. This step classifies each pixel that belongs to a particular label. This classification does not change for different instances of the same object in the structure. Therefore, this process helps to classify the various zones of the structure.
- **Post-processing (up-scaling images)**. At this stage, the segmented image is used as a mask and scaled up to the size of the raw image of the drone by implementing nearest-neighbor interpolation, edge-directed interpolation, and box sampling. Interpolation was used as when the image is scaled up it needs more pixels to fill up spaces between actual pixels. The result of the upscaling process shows how accurately the focus areas in the images have been segmented. Optionally, vegetation removal can be applied to these images based on the input data.
- **Extraction of ROI using segmentation results**. A segmented mask is defined as an eight-bit single-channel object, and each pixel has a zero or non-zero region value. When a mask object is added to the area of an image, only the non-zero area is visible, and any pixel values in the mask that overlap with the image are invisible. The same process was used to extract the ROI in the drone images using the segmentation mask results.
- **Generation of point cloud data (3D model)**. Photo collections are one of the most diverse sources of data for 3D modeling. However, using these image datasets for 3D modeling remains a challenge, owing to the ever-increasing volume of image data. Many open-source commercial 3D modeling solutions have been developed; a well-known solution is COLMAP. This software is a general-purpose SfM and MVS pipeline. It offers a wide range of features for reconstructing ordered and unordered image collections. Traditionally, image-based 3D reconstruction from images first recovers a sparse representation of the scene and the camera poses of the input images using structure-from-motion. This output then serves as the input to the MVS to recover a dense representation of the scene.

The input data are a set of overlapped extracted images of the ROI of the same object taken from different angles. The result is a 3D reconstruction of the object with the reconstructed internal and external camera parameters of all the images. In this study, point cloud data were considered as the output of photogrammetry when using COLMAP.

- **Scale adjustment**. 3D object models are widely reconstructed using SfM. However, in this case, we need to estimate the scales of the 3D scenes, because SfM does not provide scales directly. A calibration object or GCP can be used to obtain the scale of the scene; however, this is sometimes insufficient. In the case of large civil infrastructures, such as dams or slopes, using control points to adjust the scale is ineffective if the size of the object is too large. The error in the rating scale increases linearly, which significantly affects the registration result. Thus, some standard targets can be used to adjust the scale, such as the lane distance and size of the protection fans.

4.1 Semantic segmentation

At the downscaling stage, the raw drone images were scaled down to save computational time. Laboratory experiments have shown that the semantic segmentation of images reduced to 25% of original images requires approximately 50% less processing time than it takes for the segmentation of original images. Fig. 12 shows the segmentation results for the thumbnails and raw images. Downscaling removed the unwanted edges—especially in areas with vegetation—but retained the underlying context of the raw images; thus, the trained model produced better results.

Semantic segmentation is a key stage in the 3D modeling method proposed in this study. In this step, the

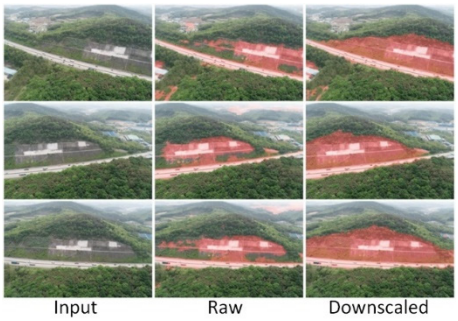


Fig. 12 Segmentation results for pre-processed images

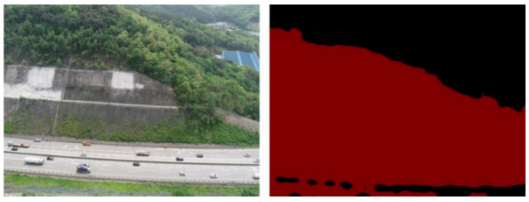


Fig. 13 Semantic segmentation output for cut-slopes

drone images are segmented using the trained model of the ROI. The trained model assigns a class to each pixel in the input image and outputs a new image of the same size, with each pixel colored according to the color of its class. Consequently, a model trained to detect cut-slope areas in drone images was used to automatically segment new cut-slope images. Fig. 13 shows the result of the cut-slope image mask that was segmented using a DeepLabv3+-based model trained on 520 images. The area of the cut slope is represented by red pixels, and areas that are not ROIs are represented by black pixels. Experiments showed that the semantic segmentation of images downscaled by 25% spent approximately 50% less time than that required for the segmentation of raw results but delivered better results overall. Regarding cut slopes, a down-scaling factor of 0.25 (25% of raw image) was used to create a low-resolution version of the high-resolution image in order to reduce computing cost. The high-frequency components of the image are lost during the downscaling, which won't affect the desired results.

Noise removal using segmentation was based on indentation images and drone video files. The implementation of the proposed 3D modeling procedure required a trained model that automatically labeled new images of civil infrastructures. The DeepLabV3+ image segmentation architecture was used in this study. DeepLabv3+ has an mIoU score of 82.1% in the PascalVOC and Cityscapes datasets; therefore, it was selected to be used in this process.

The training process includes the following steps:

- Data collection: A total of 520 cut-slope images were labeled using LabelMe for the neural network (NN) training process. These images contained all five main classes of the cut slope: slope area, fence, road, protection walls, and drainage systems.
- Configuration of hyperparameters of DeepLabV3+: In this study, a learning rate of 0.007 was used, and the number of epochs was 20. The batch size was set according to the GPU training memory specifications. A batch size of four could be used on a GPU with specs Nvidia RTX 2080 super (8GB GDDR6, 256-bit, 15.5 Gbps, 3072 nvidia CUDA cores) with 8GB RAM.
- Separation of dataset: The cut-slope dataset should be randomly divided into test and training parts; for example, out of 520 cut-slope images, 424 were used for training, and 96 were used for evaluation at each interval.

The same approach has been applied for other civil infrastructures, such as dam structures. For a dam, all the dam structures are considered the main ROI. Fig. 14 shows the image segmentation results for a dam structure.

The results of segmentation using DeepLabV3 for the dam structure case is shown in Figs. 14 and shows the mIoU scores for the learning rates for this case.

To show network implementation, evaluation indices have been calculated as following below: Igou = 82.1, F1 score = 90.1, Recall score = 95.2, Precision score = 86.1.

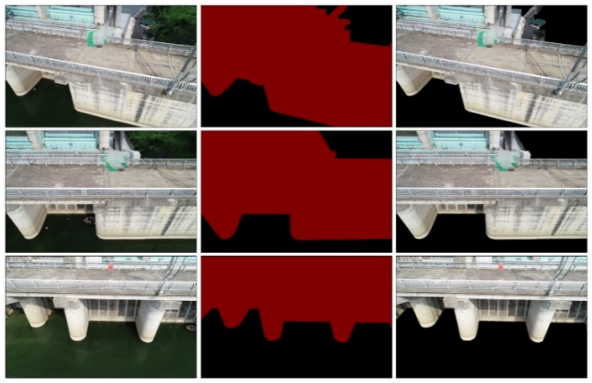


Fig. 14 Segmentation result for the dam structure

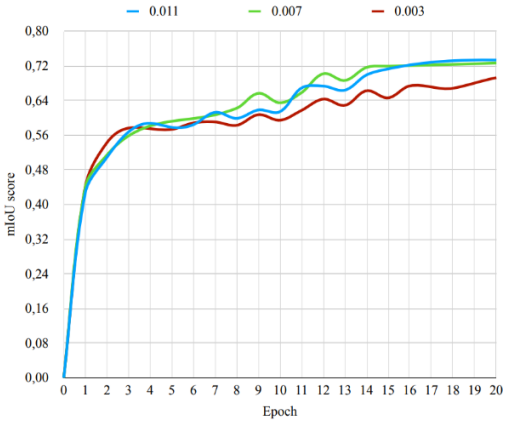


Fig. 15 mIoU scores for learning rates (0.011, 0.007, 0.003) representing the change on each of 20 epochs

4.2 3D model generation

All the drone datasets (photos) for the SfM were processed using COLMAP on a high-performance computer. COLMAP offers an automatic reconstruction feature that generates point cloud data from a set of related images. The automatic reconstruction process comprises the following steps:

- Feature detection and extraction
- Feature matching and geometric verification
- Structure and motion reconstruction

In the first stage, feature detection/extraction identifies sparse feature points in the image and describes their appearance using a numeric descriptor. COLMAP imports images and performs feature detection/extraction in one step to load the images from the disk only once. In the second step, feature matching and geometric verification determine the correspondences between the feature points in different images. In the third step, COLMAP first loads all the extracted data from the database into the memory and reconstructs the original pair of images. The scene is then gradually expanded by registering new images and triangulating new points. The results are visualized in "real-time" during the reconstruction process.

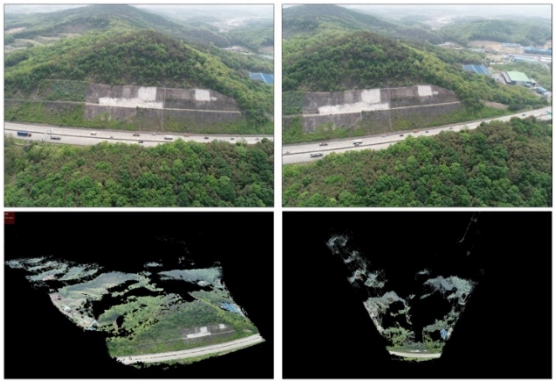


Fig. 16 Generation of a 3D model of the cut-slope using raw drone images. (top row) Depicts the raw drone images of the cut-slope from different perspectives. (bottom row) Reconstructed 3D model consists of 2.45 million points from different perspectives. The 3D point cloud model was generated from 178 images

Fig. 17 Generation of a 3D model of the cut-slope using segmented images. (top left) Raw drone image, (top right)-Segmentation mask images aimed at deriving slope ROI, (bottom row)- Reconstructed 3D model comprises essential cut-slope's two million points without noise from different perspectives

An experimental 3D point cloud model of the cut-slope object has been reconstructed using raw drone images, as shown in Fig. 16. There are raw images of the cut-slope taken by the drone (first two images from the left) and reconstructed 3D point cloud data from different perspectives (next two images from the right). The model of the cut slope resembles a slope zone and the distributed points behind it. This is because of the far background zone of the slope shown in the raw images. This occurs when photogrammetry calculates the distance and takes the key points of the image. This creates more inconvenience when 3D models were reconstructed automatically. The removal of unnecessary points from a point cloud is a time-consuming process that requires more human interaction.

Using sloped segmented images for 3D reconstruction improves the efficiency and saves processing time. The slope ROI has been extracted using the segmentation mask image, as illustrated in Fig. 17 (2nd image from the left), and it has been applied in the reconstruction of the point cloud using COLMAP. The produced result was noiseless and satisfactory, as shown in Fig. 17 (the last two images). The results were processed on a Windows PC with an AMD Ryzen 7 27000X 8- core processor, 32GB RAM, and a 3.70 GHz system.

The proposed 3D reconstruction procedure was compared with conventional reconstruction methods using the raw images of drones in terms of size, processing time, and accuracy. The image dataset of the cut slope was chosen for this comparative study. Table 1 provides an overview of the dataset used in this benchmarking study; it shows a significant reduction in the size of the raw dataset after being converted to sharded datasets. When reconstructing a 3D point cloud, segmented datasets can be used in the point-cloud generation procedure instead of the raw dataset to save computational time and memory.

Conventional 3D reconstruction methods use commercial software, such as Pix4D, with raw drone images. This section compares the proposed 3D model generation procedure with conventional methods that use Pix4D and COLMAP. 3D models were constructed by both these pieces of software, using a raw image dataset and segmented image datasets for the cut slope. As COLMAP yielded nearly the same results that Pix4D did, with denser point cloud classification, the following point clouds were generated in COLMAP.

Fig. 18 depicts comparison of two resulting point clouds from three perspectives, where a) (first row in the picture) were reconstructed by using original dataset; and b) (second row in the picture) depicts the resulting point cloud from processing of the segmented images by our proposed algorithm. Comparing the process of generation of these two results (a, b) showed that the volume of the point cloud by the proposed algorithm (b) decreased by almost 20%, and the process took 10 min less than the reconstruction process (a) with the original dataset. The first point cloud model (a) was generated using raw drone images accumulated 2.45 million points, and the processing time was 28 min. Simultaneously, the second point cloud (b) of the segmented datasets contained two million points with a processing time of 19 min.

To evaluate the quality of the 3D model generation process, comparison of the COLMAP and proposed approach was applied to dam structure. The criteria we

Table 1 Information on size of dataset

Slope side	Number of images (image resolution)	Raw dataset size	Segmented dataset size	Segmentation process time
Anseong-si slope (South Korea)	39 (3840×2160)	271.9 MB	83.7 MB	2 min 41 s

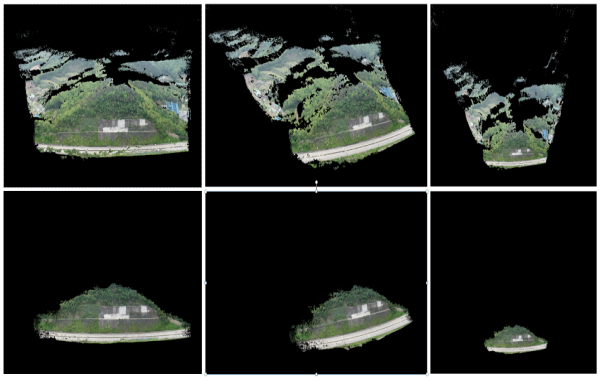
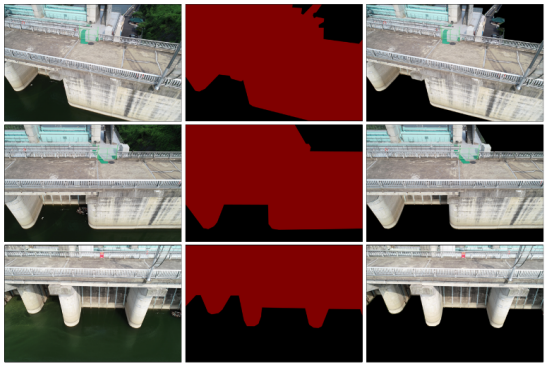
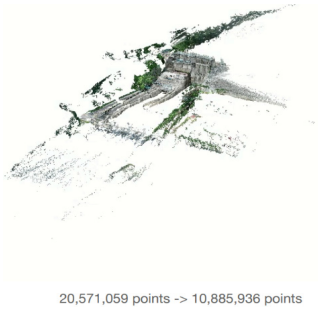


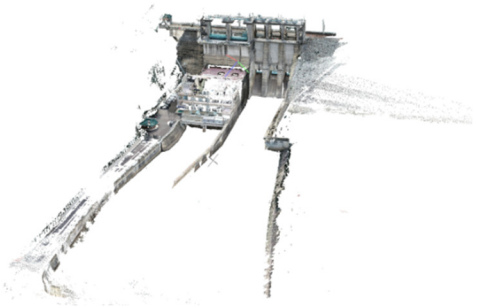
Fig. 18 Comparison of a point cloud model with raw and segmented datasets generated in COLMAP. The upper row depicts the point cloud model from various perspectives, which was generated using unprocessed drone images comprising 2.45 million points. The lower row shows the point cloud model generated from segmented images comprising two million points.



(a) Segmentation result for the dam structure



(b) The 3D point cloud model generated from raw images comprises 20 million points with processing time 2 hours



(c) The 3D point cloud model generated from segmented images comprises 10 million points with processing time 1 h 10 min

Fig. 19 Segmented images and generated dam model

considered were the presence of the noise, processing time, and the volume of the model. Regarding the dam, as shown in Fig. 19, the 3D point cloud model generated from the original dataset model shows the presence of noise (b), while the 3D model of the dam structure (c) after segmentation (a) presents a more accurate 3D model of the desired object. The comparison showed that model (b) contains 20.5 M points, and the segmented dataset model contains 10.8 M points (c); the processing times were 2 h and 1 h 10 min, respectively. The results showed that the point cloud of the segmented dam dataset consisted of 52.6% of that of the raw dataset model, and the processing time was reduced by almost 50%. The proposed method was successfully applied to automate 3D reconstruction and obtain an accurate point cloud model.

4.3 Scale adjustment

GCPs are points that a surveyor can pinpoint when taking aerial photographs: with a few known coordinates, large areas can be mapped accurately. Skipping GCPs may yield perfectly fine results, but the reconstruction may not have the correct scale, orientation, or absolute position information. GCPs or RTK (real-time kinematic) geotags can help verify the accuracy of the reconstruction. The GCPs can be easily recognized in the images. Typically, they appear as a small section of a checkerboard. The shape leaves very little ambiguity about the "point" of a GCP.

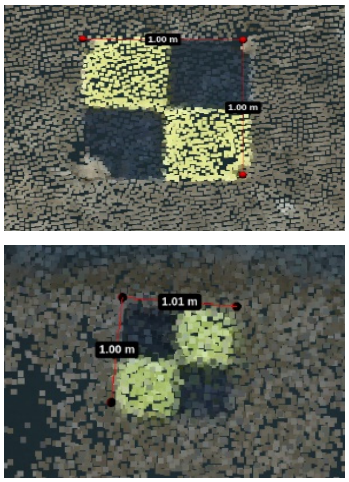
The GCPs were used to obtain the actual scale of the point cloud model. It is possible to convert the virtual size of a point cloud to its actual size by determining the size of the GCP, as shown in Fig. 20.

$$scale\ parameter = \frac{GCP\ size}{Point\ cloud\ virtual\ distance} \quad (6)$$

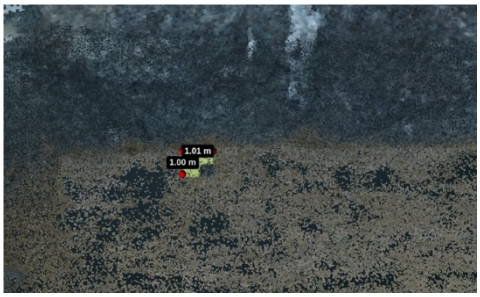
Although placing a GCP can take a long time, it takes less time than re-flying drones to improve accuracy. Many researchers have recommended using at least five GCPs. This provided more options for measuring the point cloud model after the scale parameters are calculated. Fig. 21 shows the process of measuring the length of the slope after setting the scale parameters; the result looks much more realistic.

There may be some difficulties in measuring GCPs size on point clouds, such as not dense enough points or missing the GCP while generating or capturing. In the absence of reference points, it is possible to obtain scale parameters from the known dimensions of the civil infrastructure, as shown in Fig. 22.

Fig. 22 (a) shows the real satellite image obtained using the measurement process on the KakaoMap platform. The measured result is used for calculating the scale parameters and compared with the real size of the captured infrastructure (Fig. 22(b)). In this example, the standard lane width was used to calculate the scale parameters instead of using the GCPs.



(a) GCP model 1



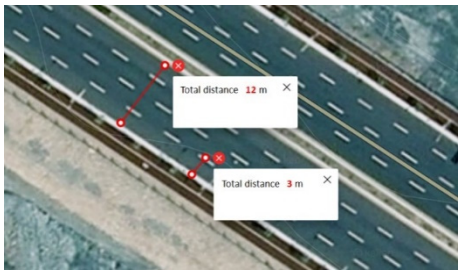
(b) Point cloud model of a slope

Fig. 20 Calculation of scale parameter from GCPs on the point cloud

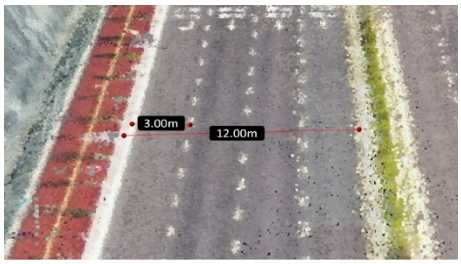
Fig. 21 Actual size of the observed objects based on the extracted scale parameter

5. Conclusions

The use of 3D models has taken the lead in the development and health inspection of smart cities and civil infrastructures. By analyzing 3D models, these tasks can be automated in the future. This study aims to improve the generation of the 3D models of civil infrastructures using deep learning and drone images. Additionally, we used this approach to analyze drone image-capturing requirements for cut slopes and dam infrastructures. These requirements comprise the guiding components for further objectives: GCPs, image-overlapping ratio, camera angle, and drone flight plan. Initially, the raw images obtained from the drones contained unwanted pixels (noise) for the 3D modeling process. In this study, we propose removing noise by implementing a deep learning approach using the semantic segmentation of raw images. The 3D construction



(a) Satellite images (KakaoMap)



(b) Generated point cloud data

Fig. 22 Comparison of satellite image measurements and generated point cloud

workflow comprises the following stages: downscaling collected drone images in accordance with the capturing requirements of the drone, semantic segmentation for noise reduction, upscaling images to their original size, extraction of the ROI, generation of a 3D model, and scale adjustment. The segmentation process uses a trained model of slopes and dams to exclude vegetation and extract ROIs from drone images by assigning one of the classes to each image pixel. Additionally, post-processing of the segmentation results and raw images was performed in the 3D model reconstruction pipeline. Using the proposed approach, a high-quality 3D point cloud model was experimentally generated in the cut-slope and dam cases; simultaneously, the size of the model was reduced and unnecessary points were removed. According to the results, the proposed method significantly improved the quality of the 3D model and reduced the model size by 20–50% after noise removal. In a benchmarking analysis, the free and open-source software COLMAP created a 3D model with qualitative characteristics of point clouds, similar to that done by commercial software. Future work will include discussions on additional applications of the proposed 3D model generation method. Further research is required to determine the cloud distance of point clouds that can be used to evaluate timely adjustments to the shapes of cut slopes and dams.

References

Agisoft Metashape Standard Edition (2021), https://www.agisoft.com/downloads/user-manuals/

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille, A.L. (2014), "Semantic image segmentation with deep convolutional nets and fully connected crfs", arXiv preprint arXiv:1412.7062. https://doi.org/10.48550/arXiv.1412.7062

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K. and Yuille,

A.L. (2017a), "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs", *IEEE Trans. Pattern Anal. Mach. Intell.*, **40**(4), 834-848. https://doi.org/10.1109/TPAMI.2017.2699184

Chen, L.C., Papandreou, G., Schroff, F. and Adam, H. (2017b), "Rethinking atrous convolution for semantic image segmentation", arXiv preprint arXiv:1706.05587. https://doi.org/10.48550/arXiv.1706.05587

Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F. and Adam, H. (2018), "Encoder-decoder with atrous separable convolution for semantic image segmentation", *Proceedings of the European Conference on Computer Vision* (*ECCV*), Munich, Germany, September.

COLMAP (2022), https://colmap.github.io/

El-Omari, S. and Moselhi, O. (2008), "Integrating 3D laser scanning and photogrammetry for progress measurement of construction work", *Automat. Constr.*, **18**(1), 1-9. https://doi.org/10.1016/j.autcon.2008.05.006

He, T., Yang, Y., Shi, Y., Liang, X., Fu, S., Xie, G., Liu, B. and Liu, Y. (2022), "Quantifying spatial distribution of interrill and rill erosion in a loess at different slopes using structure from motion (SfM) photogrammetry", *Int. Soil Water Conserv. Res.*, **10**(3), 393-406. https://doi.org/10.1016/j.iswcr.2022.01.001

Inzerillo, L., Di Mino, G. and Roberts, R. (2018), "Automation in construction image-based 3D reconstruction using traditional and UAV datasets for analysis of road pavement distress", *Autom. Constr.*, **96**, 457-469. https://doi.org/10.1016/j.autcon.2018.10.010

Irschara, A., Kaufmann, V., Klopschitz, M., Bischof, H. and Leberl, F. (2010), "Towards fully automatic photogrammetric reconstruction using digital images taken from UAVs", *Proc. Int. Soc. Photogramm. Remote Sens.*, **38**(7A), 65-70.

Jiang, Y., Bai, Y. and Han, S. (2020), "Determining ground elevations covered by vegetation on construction sites determining ground elevations covered by vegetation on construction sites using drone-based orthoimage and convolutional neural network", *J. Comput. Civil. Eng.*, **34**(6). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000930

Khaloo, A., Lattanzi, D., Cunningham, K., Dell'Andrea, R. and Riley, M. (2018), "Unmanned aerial vehicle inspection of the Placer River Trail Bridge through image-based 3D modelling", *Struct. Infrastruct. Eng.*, **14**(1), 124-136. https://doi.org/10.1080/15732479.2017.1330891

Liu, T. and Abd-Elrahman, A. (2018), "Deep convolutional neural network training enrichment using multi-view object-based analysis of Unmanned Aerial Systems imagery for wetlands classification", *ISPRS J. Photogramm. Remote Sens.*, **139**, 154-170. https://doi.org/10.1016/j.isprsjprs.2018.03.006

Liu, Z., Brigham, R., Long, E.R., Wilson, L., Frost, A., Orr, S.A. and Grau-Bové, J. (2022), "Semantic segmentation and photogrammetry of crowdsourced images to monitor historic facades", *Heritage Science*, **10**(1), 1-17. https://doi.org/10.1186/s40494-022-00664-y

Menegoni, N., Giordan, D., Perotti, C. and Tannant, D.D. (2019), "Detection and geometric characterization of rock mass discontinuities using a 3D high-resolution digital outcrop model generated from RPAS imagery – Ormea rock slope, Italy", *Eng. Geol.*, **252**, 145-163. https://doi.org/10.1016/j.enggeo.2019.02.028

Minaee, S., Boykov, Y.Y., Porikli, F., Plaza, A.J., Kehtarnavaz, N. and Terzopoulos, D. (2021), "Image segmentation using deep learning: A survey", *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**(7), 3523-3542. https://doi.org/10.1109/TPAMI.2021.3059968

Mustafa, A., Volino, M., Kim, H., Guillemaut, J.Y. and Hilton, A. (2021), "Temporally coherent general dynamic scene reconstruction", *Int. J. Comput. Vision*, **129**(1), 123-141. https://doi.org/10.48550/arXiv.1907.08195

Omar, H., Mahdjoubi, L. and Kheder, G. (2018), "Towards an automated photogrammetry-based approach for monitoring and controlling construction site activities", *Comput. Ind.*, **98**, 172-182. https://doi.org/10.1016/j.compind.2018.03.012

Pix4D (2021), https://support.pix4d.com/hc/en-us/sections/360003718992-Manual

Popescu, C., Täljsten, B., Blanksvärd, T. and Elfgren, L. (2019), "3D reconstruction of existing concrete bridges using optical methods", *Struct. Infrastruct. Eng.*, **15**(7), 912-924. https://doi.org/10.1080/15732479.2019.1594315

Rahaman, H. and Champion, E. (2019), "To 3D or not 3D: Choosing a photogrammetry workflow for cultural heritage groups", *Heritage*, **2**(3), 1835-1851. https://doi.org/10.3390/heritage2030112

Tang, S., Zhang, Y., Li, Y., Yuan, Z., Wang, Y., Zhang, X., Li, X., Zhang, Y., Guo, R. and Wang, W. (2019), "Fast and automatic reconstruction of semantically rich 3D indoor maps from low-quality RGB-D sequences", *Sensors*, **19**(3), 533. https://doi.org/10.3390/s19030533

The Cityscapes Dataset (2020), https://www.cityscapes-dataset.com/

Yücer, K., Sorkine-Hornung, A., Wang, O. and Sorkine-Hornung, O. (2016), "Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction", *ACM Transactions on Graphics* (*TOG*), **35**(3), 1-15. https://doi.org/10.1145/2876504

Zha, F., Fu, Y., Wang, P., Guo, W., Li, M., Wang, X. and Cai, H. (2020), "Semantic 3D reconstruction for robotic manipulators with an eye-in-hand vision system", *Appl. Sci.*, **10**(3), 1183. https://doi.org/10.3390/app10031183

HJ