Multi-source data based 3D digital preservation of large-scale ancient Chinese architecture: a case report

Xiang GAO¹,², Hainan CUI¹, Lingjie ZHU¹,², Tianxin SHI¹,², Shuhan SHEN¹,²*

¹. National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
* Corresponding author, shshen@nlpr.ia.ac.cn

Received: Day Month 2019 Accepted: Day Month 2019

Supported by the NSFC (National Natural Science Foundation of China) under grants 61632003 and 61873265.

Abstract Ancient Chinese architecture 3D digitalization and documentation is a challenging task due to its architectural complexity and structural delicacy. In order to generate complete and detailed models of the ancient Chinese architecture, instead of single-source data, it is better to acquire, process, and fuse multi-source data. In this paper, we describe our works on ancient Chinese architecture 3D digital preservation based on multi-source data. We first briefly introduce two ancient Chinese temples we surveyed, Foguang Temple and Nanchan Temple. Then, we report the data acquisition equipment we used and the multi-source data we acquired. Finally, we give an overview of several applications we conducted based on the acquired data, including ground and aerial image fusion, image and LiDAR (light detection and ranging) data fusion, and architectural scene surface reconstruction and semantic modeling. We believe that it is necessary to involve multi-source data for ancient Chinese architecture 3D digital preservation, and the works in this paper provide a heuristic guideline for the related research communities.

Keywords Ancient Chinese architecture; 3D digital preservation; Multi-source data acquisition; Architectural scene modeling

1 Introduction

Together with the European architecture and the Islamic architecture, the ancient Chinese architecture is one the most important components of the world architectural system. The most significant characteristic of the ancient Chinese architecture is the use of timber framework. Compared with other architectural styles, though more delicate structures can be achieved, the ancient Chinese architecture is more vulnerable to natural disasters, e.g. fire or earthquake. As a result, there is an urgent need for the preservation of the ancient Chinese architecture and one of the best means is to digitally preserve them by reconstructing their complete and detailed 3D models.

Architectural scene modeling has always been an intensive research topic in the fields of computer vision, computer graphics, and photogrammetry. Though many exciting works have been proposed, most of them
perform the modeling task with single-source data. Some methods generate the scene models from images with similar viewpoint and scale only\textsuperscript{[1–4]}, which are captured either by handheld cameras or cameras mounted on UAVs (unmanned aerial vehicles); while others try to obtain the models by using range data, e.g. RGB-D images\textsuperscript{[5–8]} obtained from Kinect or LiDAR data\textsuperscript{[9–12]} acquired by laser scanner.

However, it is hard to generate accurate and complete architectural scene models by using single-source data, especially for the ancient Chinese architecture with complicated structures. In this paper, we perform the task of ancient Chinese architecture 3D digital preservation by multi-source data acquisition, processing, and fusion. There are four types of acquired data in total, including 1) aerial images captured by a ILD (interchangeable lens digital) camera mounted on a UAV, 2) ground images captured by a DSLR (digital single lens reflex) camera mounted on a robotic camera mount, 3) laser point cloud scanned by a laser scanner, and 4) geo-coordinates of the ground control points measured by a differential GPS system. The first three types of data are used for scene modeling and complementary: the aerial images and ground images provide large-scale and close-range capturing of the scene, while the images and LiDAR data are complementary in flexibility and accuracy. The last type of data is used for image geo-reference and accuracy evaluation.

In the following of this paper, we first briefly introduce the ancient Chinese temples we surveyed. Then, we report the used data acquisition equipment and the acquired multi-source data. Finally, we give an overview of our conducted several applications based on the acquired data.

2 Scenes

The architectural scenes we surveyed in this paper are two ancient Chinese temples, named Foguang Temple and Nanchan Temple respectively (cf. Fig. 1). They are two of four existing Chinese architecture with timber structure and built in Tang Dynasty. Among them, Foguang Temple is the largest one while Nanchan Temple is the earliest one. Some details of the two temples are described in the following respectively.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{MVS_point_clouds.png}
\caption{MVS (multi-view stereo) point clouds of FGT (Foguang Temple) and NCT (Nanchan Temple), which are generated by the method \textsuperscript{[13]}. The red, green, and blue rectangles in the left figure denote the GEH (Great East Hall), MJH (Manjusri Hall), and GRH (Garan Hall) of FGT. The red rectangle in the right figure denotes the GBH (Great Buddha Hall) of NCT.}
\end{figure}
2.1 Fuguang Temple

FGT (Foguang Temple) is a Buddhist temple located in Wutai County, Shanxi Province of China. Its covering area is about 34000m². It mainly contains three halls, named Great East Hall, Manjusri Hall, and Garan Hall, respectively.

2.1.1 Great East Hall
Dating from 857 of the Tang Dynasty, the GEH (Great East Hall) is the third oldest dated but the largest wooden building in China. The hall is located on the far-east side of the temple. It is a single storey structure measuring seven bays by four, and is supported by inner and outer sets of columns. On the top of each column is a complicated set of brackets containing seven different bracket types. Inside the hall are thirty-six sculptures, as well as murals on each wall that date from the Tang Dynasty and later periods.

2.1.2 Manjusri Hall
On the north side of the temple courtyard is the MJH (Manjusri Hall). It was constructed in 1137 during the Jin Dynasty and is roughly the same size as the GEH. It features a single-eave hip gable roof. The interior of the hall has only four support pillars. In order to support the large roof, diagonal beams are used. On each of the four walls are murals of arhats painted in 1429 during the Ming dynasty.

2.1.3 Garan Hall
The GRH (Garan Hall) is located in the southwest corner of FGT. It was first built in 1628 of the Ming Dynasty and rebuilt in 1661 of the Qing Dynasty, measuring three bays in width. The sculptures of eighteen Gran Gods are in the GRH.

2.2 Nanchan Temple

NCT (Nanchang Temple) is a Buddhist temple which is also located Wutai County, Shanxi Province of China. Its covering area is about 3100m². In the NCT, there is only one main hall, named Great Buddha Hall.

2.2.1 Great Buddha Hall
Built in 782 during Tang Dynasty, the GBH (Great Buddha Hall) is the oldest preserved timber building extant of China. It is a three bay square hall. Not only is the GBH an important architecture, but it also contains an original set of artistically-important Tang sculptures dating from the period of its construction. Seventeen sculptures share the interior space of the hall with a small stone pagoda.

3 Equipment

The equipment for data acquisition in this paper are divided into four categories according to the data types, i.e. equipment for 1) aerial images, 2) ground images, 3) LiDAR data acquisition, and 4) ground control points (cf. Fig. 2). In the following, we introduce what equipment we used and describe each one of them briefly.

1 https://en.wikipedia.org/wiki/Foguang_Temple
Figure 2 Equipment for multi-source data acquisition in this paper.

3.1 Equipment for aerial image acquisition

For aerial image acquisition, we use an ILD camera, Sony NEX-5R, mounted on a UAV, Microdrones Md4-1000.

3.1.1 Sony NEX-5R

The Sony NEX-5R is an ILD camera with a 16.1 effective megapixel CMOS (complementary metal oxide semiconductor) sensor. It has similar imaging quality but much lighter weight compared with standard DSLR camera. The above features make the Sony NEX-5R more suitable for aerial image capturing.

3.1.2 Microdrones Md4-1000

The Microdrones Md4-1000 system is a leading VTOL AUMAV (vertical take-off and landing, autonomous unmanned micro aerial vehicle). The drone body and camera mount are made of carbon fiber material which is lighter in weight and higher in strength.

3.2 Equipment for ground image acquisition

For ground image acquisition, we use a DSLR camera, Canon EOS 5D Mark III, mounted on a robotic camera mount, GigaPan Epic Pro.

3.2.1 Canon EOS 5D Mark III

The Canon EOS 5D Mark III is one of the most famous DSLR cameras. It is equipped with a 22.3 megapixel full-frame CMOS sensor and has excellent imaging quality under various environments. As a result, the Canon EOS 5D Mark III is a suitable choice for scene capturing from ground viewpoint.

---

3 https://www.sony.com/
4 https://www.microdrones.com/
5 https://www.usa.canon.com/
3.2.2 GigaPan Epic Pro

The GigaPan Epic Pro is a robotic camera mount that could capture HD, gigapixel photos using almost any digital camera. By setting the upper left and lower right corners of the panorama desired, the GigaPan Epic Pro works out how many photos the camera need to take, and then automatically organizes them.

3.3 Equipment for LiDAR data acquisition

For LiDAR data acquisition, we use a laser scanner, Leica ScanStation P30.

3.3.1 Leica ScanStation P30

The Leica ScanStation P30 delivers high quality 3D data and HDR (high dynamic range) imaging at an extremely fast scan rate of 1 million points per second and at ranges of up to 270m with extremely high accuracy. For example, its 3D position accuracy is 3mm at 50m and 6mm at 100m, respectively.

3.4 Equipment for ground control point measurement

For ground control point measurement, we use a differential GPS system, Hi-Target V30 GNSS RTK.

3.4.1 Hi-Target V30 GNSS RTK

The V30 GNSS RTK possesses outstanding positioning performance. For example, its horizontal positioning accuracy and vertical positioning accuracy in the high-precision static situation are 2.5mm + 0.1ppm RMS (root-mean-square) and 3.5mm + 0.4ppm RMS, respectively.

4 Data

In this section, we introduce the multi-source data we acquired in the scenes described in Sec. 2 with the equipment described in Sec. 3. The acquired multi-source data consists of aerial images, ground images, LiDAR data, and ground control points.

4.1 Aerial images

We manually fly the Microdrones Md4-1000 in FGT and NCT and trigger the Sony NEX-5R shutter to capture aerial images. The images are captured with five flight paths, one for nadir images and the other four for 45° oblique images. These images are with the resolution of 4912 × 3264. We took 1596 aerial images for FGT and 772 images for NCT. The aerial image examples are shown in Fig. 3. In addition, the SfM (structure from motion) point clouds and camera poses of the FGT and NCT aerial images are computed by the method [1], which are shown in Fig. 4.

---

6 http://www.gigapan.com/
7 http://www.leica-geosystems.com/
Figure 3 Aerial image examples of FGT (Foguang Temple) and NCT (Nanchang Temple).

Figure 4 Aerial SfM (structure from motion) point clouds of FGT (Foguang Temple) and NCT (Nanchang Temple) generated by the method [1]. The red cones denote the poses of aerial cameras.

4.2 Ground images

Figure 5 Ground image examples of FGT (Foguang Temple), including outdoor image examples of FGT; indoor image examples of GEH (Great East Hall), MJH (Manjusri Hall), and GRH (Garan Hall).
We mount the Canon EOS 5D Mark III on the Gigapan Epic Pro and take ground images station by station. The Gigapan Epic Pro is set to capture images with pitch range $-40^\circ \sim 40^\circ$, step $20^\circ$ and yaw range $0^\circ \sim 320^\circ$, step $40^\circ$, which means 45 ground images are captured each station. The captured ground images are with the resolution of $5760 \times 3840$. There are 155, 55, 32, and 6 image capturing stations for outdoor scenes of FGT and indoor scenes of GEH, MJH, and GRH, respectively. In addition, there are 62 and 19 image capturing stations for outdoor scenes of NCT and indoor scenes of GBH, respectively. The ground image examples of FGT and NCT are shown in Fig. 5 and Fig. 6, respectively. In addition, the SfM point clouds and camera poses of the FGT and NCT ground images are computed by the method [1], which are shown in Fig. 7.

![Outdoor image examples of NCT](image1.png)

![Indoor image examples of GBH](image2.png)

Figure 6 Ground image examples of NCT (Nanchang Temple), including outdoor image examples of NCT; indoor image examples of GBH (Great Buddha Hall).

![Outdoor SfM point cloud of FGT](image3.png)

![Outdoor SfM point cloud of NCT](image4.png)

![Indoor SfM point cloud of GEH](image5.png)

![Indoor SfM point cloud of MJH](image6.png)

![Indoor SfM point cloud of GRH](image7.png)

![Indoor SfM point cloud of GBH](image8.png)

Figure 7 Ground SfM (structure from motion) point clouds of FGT (Foguang Temple) and NCT (Nanchang Temple) generated by the method [1], including outdoor SfM point clouds of FGT and NCT; indoor SfM point clouds of GEH (Great East Hall), MJH (Manjusri Hall), GRH (Garan Hall), and GBH (Great Buddha Hall). The red cones denote the poses of aerial cameras.

4.3 LiDAR data
Figure 8 Laser point cloud examples of FGT (Foguang Temple) and NCT (Nanchan Temple), including outdoor laser point cloud examples of FGT and NCT; indoor laser point cloud examples of GEH (Great East Hall), MJH (Manjusri Hall), GRH (Garan Hall), and GBH (Great Buddha Hall).

We use Leica ScanStation P30 to acquire LiDAR data of FGT and NCT. We first determine the laser scanning locations and then perform scanning. There are 39, 35, 16, and 3 laser scanning stations for outdoor scenes of FGT and indoor scenes of GEH, MJH, and GRH, respectively. In addition, there are 12 and 8 laser scanning stations for outdoor scenes of NCT and indoor scenes of GBH, respectively. For each station, we can obtain about 100 million high accuracy laser points with RGB information. The laser point cloud examples of FGT and NCT are shown in Fig. 8. In addition, the locations of outdoor laser scanning stations of FGT and NCT are shown in Fig. 9.

Figure 9 Locations of outdoor laser scanning stations and GCP (ground control point) examples for aerial cameras of FGT (Foguang Temple) and NCT (Nanchan Temple), which are marked by “★” and “+”, respectively.

4.4 Ground control points

The geo-coordinates of the GCPs (ground control points) are measured by the V30 GNSS RTK system. The GCPs have two usages in this paper: 1) they are used to geo-refer the (aerial and ground) images and 2) served as ground truths to evaluate the calibration results of the (aerial and ground) cameras. There are two
types of GCPs according to the camera types: 1) GCPs for aerial cameras. These GCPs are manually selected in the scenes and marked in the aerial images, thus they are usually obvious corners. There are 53 and 33 GCPs of this kind for FGT and NCT, respectively, and Fig. 9 and Fig. 10 give several examples of them. 2) GCPs for outdoor ground cameras. These GCPs are located at the outdoor image capturing stations to record their accurate geo-coordinates. As a result, there are 155 and 62 GCPs of this kind for FGT and NCT, respectively, which are the same as the number of outdoor image capturing stations.

5 Applications

Based on the acquired multi-source data, we mainly conduct four types of applications, including 1) aerial and ground image fusion\textsuperscript{[13]}, 2) image and LiDAR data fusion\textsuperscript{[26]}, 3) surface reconstruction\textsuperscript{[27]}, and 4) semantic modeling\textsuperscript{[36]}. They are introduced in the following.

5.1 Aerial and ground image fusion

In order to reconstruct a complete 3D digital model of ancient Chinese architecture that captures details of complex structures, \textit{e.g.} cornices and brackets, usually two different sources of images, aerial and ground, are involved for large-scale and close-range scene capturing. When using both aerial and ground images, a common practice is to carry out the reconstruction separately to generate aerial and ground point clouds at first and then fuse them afterwards. Considering the noisy nature of reconstructed 3D point clouds from image collections and the loss of rich textural and contextual information of 2D images in 3D point clouds, it is preferable to fuse the point clouds via 2D image feature point matching rather than by direct 3D point cloud registration, \textit{e.g.} ICP (iterative closest point)\textsuperscript{[14]}. In order to fuse the aerial and ground images for complete scene model reconstruction, two understanding issues should be specially addressed: 1) how to match the aerial and ground images with substantial variations in viewpoint and scale; and 2) how to fuse the aerial and ground point clouds with drift phenomena and notable differences in noise level, density and accuracy.

To deal with the aerial and ground image matching problem, in [13], the ground image is warped to the viewpoint of the aerial image, by which the differences in viewpoint and scale between these two kinds of images are eliminated. Unlike the method in [15], which synthesizes the aerial-view image by leveraging
the spatially discrete ground MVS point cloud, the image synthesis method in [13] resorts to the spatially continuous ground sparse mesh, which is reconstructed from the ground SfM point cloud. For a pair of aerial and ground images, each spatial facet in their co-visible ground sparse mesh induces a local homography between them. The aerial-view image is synthesized by warping the ground image to the aerial one using the induced homographies. Note that the above image synthesis method is free from the time-consuming MVS procedure and the resultant synthetic images would not suffer from missing pixels in the co-visible regions of aerial and ground image pairs.

Figure 11 An image feature matching example of a pair of aerial and ground images by the method in [13]. The first row is the matching result between the co-visible regions of the aerial and synthetic images, where the blue segments denote the point matches. The second row is the original aerial and ground image matching pair, where the black rectangles denote the co-visible regions for image matching.

Figure 12 A cross-view track linking example by the method in [13]. The first row contains three aerial and three ground image patches, where the blue segment denotes the linked track across views. The second row contains original aerial and ground images, where the black rectangles denote the image patches in the first row.

After image synthesis, the synthetic image is matched with the target aerial image by SIFT (scale invariant feature transform)[16] feature point extraction and matching. Then, instead of filtering out the inevitable point match outliers by NNDR (nearest neighbor distance ratio) test[17], which is prone to discarding true positives, it is done in [13] by the following two techniques: 1) a consistency check of the
feature scales and principal orientations between the point matches and 2) an affine transformation verification of the feature locations between the point matches. Note that, unlike the commonly used fundamental matrix based outlier filtering scheme which provides point-to-line constraint, the affinity based one in [13] provides point-to-point constraint, thus it is more effective for outlier filtering. Fig. 11 gives an image feature matching example of a pair of aerial and ground images.

Figure 13 Aerial and ground SfM (structure from motion) point cloud fusion results of FGT (Foguang Temple) and NCT (Nanchan Temple) by the method in [13]. The first row: result of NCT, the second row: result of FGT. From left to right: ground and aerial image examples, ground SfM point clouds, aerial SfM point clouds, and fused SfM point clouds.

To tackle the aerial and ground point cloud fusion issue, rather than aligning the point clouds by estimating a similarity transformation[18] between them with RANSAC (random sample consensus)[19], which is done in [15, 20, 21], the point clouds are fused together by a global BA (bundle adjustment)[22] to deal with the possible scene drift phenomenon. To achieve that, in [13], the obtained aerial and ground point matches are linked to the original aerial tracks at first. Fig. 12 gives a cross-view track linking example. Then, a global BA is performed to fuse the aerial and ground SfM point clouds with the augmented aerial tracks and the original ground tracks. Fig. 13 shows the aerial and ground SfM point cloud fusion results of FGT and NCT.

5.2 Image and LiDAR data fusion

There are two key issues in reconstructing large-scale architectural scenes: accuracy and completeness. Though many existing methods focus on the issue of reconstruction accuracy, they pay less attention to the reconstruction completeness. When the architectural scene is complicated, e.g. ancient Chinese architecture, the reconstruction completeness of the common pipelines is hard to guarantee. In order to reconstruct accurate and complete 3D models (point clouds or surface meshes) of the large-scale and complicated architectural scenes, both global structures and local details of the scenes need to be surveyed. Currently, there are two frequently used surveying ways for scene reconstruction, image based[1–4] and laser scanning based[9–12] methods. These two approaches are complementary in flexibility and accuracy: the image based
reconstruction methods are convenient and flexible, but they heavily depend on several external factors, e.g. illumination variation, textural richness, and structural complexity; while the laser scanning based reconstruction methods possess high accuracy and are robust to adverse conditions, but they are high-cost and time-consuming.

In order to generate a complete scene reconstruction by fusing images and LiDAR data, a straightforward way is to treat images and LiDAR data equally. Specifically, architectural scene models are obtained from these two kinds of data respectively at first and fused together by GCPs\(^{[23]}\) or using ICP algorithm\(^{[24,25]}\) afterwards. However, this is non-trivial because the point clouds generated from images and laser scans have significant differences in density, accuracy, completeness, etc. which results in inevitable registration errors. In addition, the laser scanning locations need to be carefully selected to guarantee the scanning overlap for their self-registration.

To deal with the above issues, we propose a more effective data collection and scene reconstruction pipeline in \([26]\), which takes both the data collection efficiency, and the reconstruction accuracy and completeness into consideration. Our pipeline uses images as primacy to completely cover the scene, and uses laser scans as supplement to deal with low textured, low lighting, or complicated structured regions. Similar to \([13]\), in \([26]\), images and LiDAR data are fused by 2D image feature point matching between captured images and images synthesized from LiDAR data, instead of 3D point cloud registration.

In \([26]\), we first obtain fused SfM point cloud from the captured aerial and (outdoor and indoor) ground images. To achieve this, both point matches between aerial and ground images and between outdoor and indoor images are required. However, obtaining these two kinds of point matches are both non-trivial, due to 1) the large viewpoint and scale differences between the aerial and ground images, and 2) the limited view overlapping between the outdoor and indoor images. In \([26]\), we generate SfM point clouds from aerial, outdoor, and indoor images individually at first and then fuse them with the help of the cross-view point matches. The aerial and ground point matches are obtained by the method in \([13]\), while the outdoor and indoor point matches are obtained by matching the outdoor and indoor images near the door.

After that, the aerial-view and ground-view synthetic images are generated from the laser point clouds and are matched with the captured ones to obtain cross-domain correspondences. Fig. 14 and Fig. 15 give image feature matching examples of a pair of synthetic aerial-view and captured aerial images and a pair of synthetic ground-view and captured ground images, respectively. Based on the cross-domain 2D point matches, the images and LiDAR data are fused in a coarse-to-fine scheme. The laser point cloud of each scanning station is coarsely registered to the fused SfM point cloud individually by a similarity transformation\(^{[18]}\) between them, which is estimated using RANSAC\(^{[19]}\). The 3D point correspondences for similarity transformation estimation are converted from the obtained cross-domain 2D point matches. Then, the camera poses of the captured images, the spatial coordinates of the SfM point cloud, and the alignments of the laser scans are jointly optimized by a generalized BA to finely merge the images and LiDAR data. The BA procedure in \([26]\) is called a generalized one because the camera poses and the laser scan alignments are simultaneously optimized by minimizing both 2D-3D reprojection errors and 3D-3D space errors. Fig. 16 shows the SfM and laser point cloud fusion results of FGT and NCT.
Figure 14 An image feature matching example of a pair of synthetic and aerial images by the method in [26]. The first row is enlarged synthetic and aerial image pair of the green rectangles in the second row to illustrate the point matches, which are denoted by the blue segments. The second row is the original synthetic and aerial image pair.

Figure 15 An image feature matching example of a pair of synthetic and ground images by the method in [26], where the blue segments denote the point matches.

Figure 16 SfM (structure from motion) and laser point cloud fusion results of FGT (Foguang Temple) and NCT (Nanchan Temple) by the method in [26]. The first row: result of NCT, the second row: result of FGT. From left to right: (aerial-outdoor-indoor) SfM point clouds, (outdoor-indoor) laser point clouds, and fused SfM and laser point clouds (red for laser point clouds, green for aerial SfM point clouds, and blue for ground SfM point clouds).
5.3 Surface reconstruction

Figure 17 Surface reconstruction result (right) of FGT (Foguang Temple) from redundant and noisy MVS (multi-view stereo) point cloud (left) by the method in [27].

Though tremendous progress in the community of image based architectural scene reconstruction has been made in recent years, when it comes to large-scale scenes with multi-scale objects, current reconstruction methods have some problems with the completeness and accuracy, especially when concerning scene details.

Scene details such as small-scale objects and object edges are an essential part of scene surfaces. Fig. 17 shows an example of preserving scene details in reconstructing FGT. In general, representing scene details, e.g. the brackets in Fig. 17, in cultural heritage digitalization projects is among the most important tasks. The point cloud representation is often redundant and noisy, while the mesh representation is concise but it sometimes loses some information. Therefore, preserving scene details in reconstructing multi-scale scenes has been a difficult problem in surface reconstruction. The existing surface reconstruction methods [28–31] either ignore the scene details or rely on further refinement to restore them. Firstly, this is because, compared with noise, the supportive points in such part of the scene are sparse, making it difficult to distinguish true surface points from false ones. Secondly, the visibility models and associated parameters employed in existing methods are not particularly suitable for large scale ranges, where scene details are usually compromised for overall accuracy and completeness. While the first case seems to be unsolvable due to the lack of sufficient information, we focus on the second one in [27].

Figure 18 Surface reconstruction without and with the likelihood energy proposed in [27]. From left to right: MVS (multi-view stereo) point cloud with heavy noise, reconstructed meshes without and with the likelihood energy.

In many previous surface reconstruction methods [28–30], visibility information that records a 3D point is seen by the views used to help to generate accurate surface meshes. To use the visibility information,
assumptions of the visibility model are made so that the space between camera center and the 3D point is free-space and the space behind the point along the line of sight is full-space. However, the above visibility model has two shortcomings: 1) the points are often contaminated with noise and 2) the full-space scales are often hard to determinate. To deal with these issues, the main works and contributions of our method in [27] are three-fold, which are listed in the following. 1) To preserve scene details without decreasing the noise filtering ability, we propose a new visibility model with error tolerance and adaptive end weights. 2) We also introduce a new likelihood energy representing the punishment of wrongly classifying a part of space as free-space or full-space, which helps to improve the ability of the proposed method to efficiently filter noise (cf. Fig. 18). 3) Moreover, we further improve the performance of the proposed method with the dense visibility technique, which helps to keep the object edge sharp (cf. Fig. 19).

![Figure 19 Surface reconstruction without and with the dense visibility technique proposed in [27]. From left to right: original image and its depth map, reconstructed meshes without and with the dense visibility technique.](image)

### 5.4 Semantic modeling

3D semantic modeling from images has gained its popularity in recent years. Its goal is to obtain both 3D structure and semantic knowledge of a scene. 3D semantic models help humans and automatic systems know “what” is “where” in a specific scene, which is a stated goal of computer vision and has a variety of applications in fields like automatic piloting, augmented reality, and service robotics. Over the last decades, tremendous progress has been made in the field of 3D geometric reconstruction, which enables us to reconstruct large-scale scenes at a high level of details. At the same time, deep learning techniques have led to a huge boost in 2D image understanding, such as semantic segmentation and instance recognition. Thus, combining deep learning and geometry reconstruction to acquire 3D semantic models interests more and more researchers nowadays. Generally, there are two ways to achieve this goal, the first is to jointly optimize the 3D structure and semantic meaning of the scene[^30,32,33], and the second is to assign semantic labels to the estimated 3D structure[^34–36]. Our work in [37] falls into the second category, i.e. we focus on labeling existing 3D geometry models, especially fine-level labeling of large-scale mesh models.

With the help of the state-of-the-art SfM[^1,38] and MVS[^2,31] algorithms, detailed 3D model could be reconstructed from hundreds and thousands of images. A straightforward way to label this model is to annotate each facet directly. However, this process is quite cumbersome because there is no effective tool for manual annotation in 3D space, and current deep learning based labeling pipeline like [39, 40] cannot
deal with large-scale 3D models. Thus, a feasible method for large-scale 3D model labeling is to firstly perform pixel-wise semantic segmentation on 2D images and then back-project these labels into 3D space using the calibrated camera parameters and fuse them together. Apparently, in this way the quality of the 3D semantic labeling highly depends on that of the 2D semantic segmentation. Current 2D semantic segmentation methods tend to fine-tune a pre-trained CNN (convolutional neural network) within the transfer learning framework, but still require a large number of manually annotated images for cross-domain datasets. However, in specialized domains such as fine-level labeling of ancient Chinese architecture, only experts with special knowledge and skills can annotate them reliably. Therefore, reducing the cost of annotation is meaningful. In [37], we propose a novel method that could dramatically reduce the annotation cost by integrating AL (active learning) into the fine-tuning process. AL is an established way to reduce the labeling workload by iteratively choosing images for annotation to train the classifier for better performance.

![Figure 20 Pipeline of the method proposed in [37].](image)

![Figure 21 Semantic modeling results of FGT (Foguang Temple) and NCT (Nanchan Temple) by the method in [37]. The first row: result of NCT, the second row: result of FGT.](image)

In [37], we start to fine-tune a CNN for image semantic segmentation with limited number of annotated images, and use it to segment all other unannotated images. Then, all predicted image labels are back-projected into 3D space and fused on the 3D model using MRF (Markov random field). Since the 3D
semantic model takes both 2D image segmentation and 3D geometry into consideration, it could be used as a reliable intermediate to select most worthy image candidates for annotation and then proceed the next fine-tune iteration. This training-fusion-selection process continues until the label configuration of the model becomes steady. Fig. 20 shows the pipeline of our method proposed in [37] and Fig. 21 shows the semantic modeling results of FGT and NCT.

6 Conclusions

In this paper, we give a report on our works of 3D digital preservation of large-scale ancient Chinese architecture based on multi-source data. We first introduce two famous ancient Chinese temples we surveyed, FGT and NCT. Then, we briefly introduce the data acquisition equipment we used, including: 1) Sony NEX-5R and Microdrones Md4-1000 for aerial images, 2) Canon EOS 5D Mark III and GigaPan Epic Pro for ground images, 3) Leica ScanStation P30 for LiDAR data, and 4) Hi-Target V30 GNSS RTK for GCPs. Subsequently, we report the multi-source data acquired by the above equipment and show several examples of them. Finally, we give an overview of several applications we conducted based on the multi-source data, including ground and aerial image fusion[13], image and LiDAR data fusion[26], and architectural scene surface reconstruction[27] and semantic modeling[37]. We believe that involving multi-source data is a more effective way for ancient Chinese architecture 3D digital preservation, and the works performed in this paper could be served as a heuristic guideline for the related research communities.

7 References

   DOI: 10.1109/CVPR.2017.257
   DOI: 10.1109/TIP.2013.2237921
   DOI: 10.1007/s11263-017-1017-7
   DOI: 10.1007/978-3-030-01252-6_38
   DOI: 10.1109/CVPR.2015.7299195
   DOI: 10.1145/3072959.3054739


M. Li, P. Wonka, and L. Nan. Manhattan-world urban reconstruction from point clouds. In European Conference on Computer Vision (ECCV), pp. 54–69, 2016. DOI: 10.1007/978-3-319-46493-0_4


DOI: 10.1109/3DV.2018.00066

DOI: 10.1109/CVPR.2016.445

DOI: 10.1109/CVPR.2017.16