FUSED 3D TRANSPARENT VISUALIZATION FOR LARGE-SCALE CULTURAL HERITAGE USING DEEP LEARNING-BASED MONOCULAR RECONSTRUCTION

Jiao Pan[1,*], Liang Li[2], Hiroshi Yamaguchi[3], Kyoko Hasegawa[2], Fadjar I.Thufail[4], Brahmantara[5], Satoshi Tanaka[2]

[1]Graduate School of Information Science and Engineering, Ritsumeikan University, Japan - gr0342ir@ed.ritsumei.ac.jp
[2]College of Information Science and Engineering, Ritsumeikan University, Japan
[3]Nara National Research Institute for Cultural Properties, Japan
[4]Research Center for Area Studies (P2W), Indonesian Institute of Sciences (LIPI), Jakarta, Indonesia
[5]Borobudur Conservation Office, Magelang, Indonesia

KEY WORDS: digital archives, transparent visualization, 3D reconstruction, photogrammetry, deep learning, cultural heritage

ABSTRACT:

This paper proposes a fused 3D transparent visualization method with the aim of achieving see-through imaging of large-scale cultural heritage by combining photogrammetry point cloud data and 3D reconstructed models. 3D reconstructed models are efficiently reconstructed from a single monocular photo using deep learning. It is demonstrated that the proposed method can be widely applied, particularly to instances of incomplete cultural heritages. In this study, the proposed method is applied to a typical example, the Borobudur temple in Indonesia. The Borobudur temple possesses the most complete collection of Buddhist reliefs. However, some parts of the Borobudur reliefs have been hidden by stone walls and became not visible following the reinforcements during the Dutch rule. Today, only gray-scale monocular photos of those hidden parts are displayed in the Borobudur Museum. In this paper, the visible parts of the temple are first digitized into point cloud data by photogrammetry scanning. For the hidden parts, a 3D reconstruction method based on deep learning is proposed to reconstruct the invisible parts into point cloud data directly from single monocular photos from the museum. The proposed 3D reconstruction method achieves 95% accuracy of the reconstructed point cloud on average. With the point cloud data of both the visible parts and the hidden parts, the proposed transparent visualization method called the stochastic point-based rendering is applied to achieve a fused 3D transparent visualization of the valuable temple.

1. INTRODUCTION

Tangible cultural heritage sites can be damaged or destroyed accidentally, deliberately, or by a natural disaster, which is a huge loss to the civilization. With the rapid development of laser scanning and photogrammetry scanning techniques, these problems can be solved by establishing digital archives of tangible cultural heritage sites. The visualization of digital archives is increasingly important for the preservation and analysis of cultural heritage sites. Many meaningful applications such as walk-through displays, computer-aided design, geographic information systems, and virtual reality applications can be implemented based on the visualization of digital archives. All these methods require an efficient and accurate digitizing method of cultural heritage sites.

Nowadays, it is efficient to acquire and preserve digital data of extant cultural heritages using 3D scanning technologies and 3D modeling tools. However, there are a lot of cases where many cultural heritage sites no longer exist or are partially damaged. In this situation, 3D scanning technologies are insufficient for establishing a complete digital archive of their original appearance. If adequate photos from different directions remain available, 3D models of the damaged parts can be reconstructed by multiple image-based methods. However, in a lot of cases, there is only a single monocular photo per object. Hence, for this kind of cultural heritage sites, a method that can perform 3D reconstruction from a single image is urgently required. In addition, cultural heritage sites are generally broad in their scale. Therefore, the efficiency of the reconstruction method is extremely important. To address this issue, this study proposes a method based on deep learning to efficiently reconstruct cultural heritage sites from a single monocular photo. The proposed method is successfully applied to a typical example, the Borobudur temple, and its results are promising.

The Borobudur temple in Indonesia is a UNESCO World Heritage Site and the largest Buddhist temple in the world. The temple consists of six square and three circular stacked platforms topped by a central dome. This temple comprises approximately 2,672 individual bas-reliefs (sculptural reliefs in which forms extend only slightly from the background) containing 1,460 narrative and 1,212 decorative panels distributed on six square platforms. These reliefs can be divided into five sections based on different independent stories they tell. The section that needs to be reconstructed is called Karmavibhangga and it consists of 160 reliefs distributed on the first square platform. This temple of high cultural value has been restored and its foot encasement was reinstalled owing to safety concerns. During the restoration, the Karmavibhangga reliefs on the first square floor were covered by stones. Since then, the first square floor has been called the hidden foot and the Karmavibhangga reliefs have been hidden from common visitors. Today, only the southeast corner with four Karmavibhangga reliefs of the hidden foot is revealed. For the hidden 156 reliefs, only gray-scale photos taken in 1890 remain. For each relief, there is only one photo taken right in front of the reliefs (see Fig. 1).

This paper proposes a method for visualizing the original appearance of this valuable cultural heritage site before its parts were covered. It bases on a deep-learning-based 3D reconstruction method and a transparent visualization method called the stochastic point-based rendering (SPBR). For the visible parts of the Borobudur temple, photogrammetry scanning is used to

[*] Corresponding author

https://doi.org/10.5194/isprs-annals-V-2-2020-989-2020

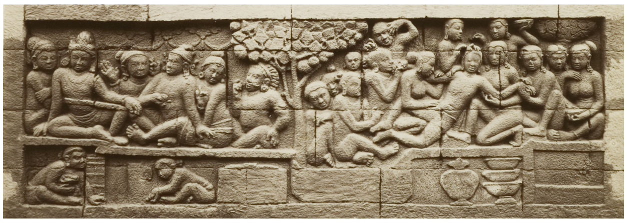


Figure 1. An example of old photos taken in 1890

digitize surfaces into point cloud data. For the hidden parts, the hidden reliefs are reconstructed into point clouds from their monocular gray-scale photos. Then the photogrammetry data of the visible parts and reconstructed data of the hidden reliefs are combined after a coordinate transformation. After the stochastic point-based rendering is applied, the result provides see-through imaging of the entire temple as well as the appearance of reliefs which are supposed to be covered by stones, such combination is defined as fused transparent visualization in this paper. With fused transparent visualization, the hidden reliefs of Karmavibhangga can be clearly seen through the stone walls.

The proposed 3D reconstruction method is based on a depth estimation neural network that maps intensity or color measurements to depth values. Compared to other methods that need hand-crafted features or manual annotation, the proposed method is more efficient and suitable for large-scale cultural heritage sites. Once the model is properly trained, the hidden reliefs can be reconstructed from a single monocular photo within a few seconds. Supervised learning is used to train the depth estimation neural network with a training data set obtained from the photogrammetry scanning data of the visible reliefs in Borobudur. With the information of 3D coordinator in point clouds, the monocular images and corresponding depth maps of the visible parts of the Borobudur reliefs can be created. Herein, ten visible reliefs containing four reliefs of Karmavibhangga and six reliefs from other sections are used as our training data set. After the model is trained properly, a depth map can be predicted by the neural network. While the study (Pan et al., 2018) suffered from low output resolution and overfitting during training and testing, the model used in the present work is inspired by the residual network (ResNet) with higher output resolution, which should solve these issues. As a result, the accuracy of the reconstructed model reaches 95%, which is 5% more compared to the result of (Pan et al., 2018).

Furthermore, to provide a good understanding of the Borobudur temple's complex internal structure as well as the original appearance of Karmavibhangga reliefs, this study applies transparent visualization to the point clouds. As the Borobudur temple is a large cultural heritage site, the total point number of 3D points in the photogrammetry point cloud reaches about 10^{14}. For conventional transparent visualization methods, depth sorting is performed to achieve the correct depth feel. Every rendering primitive is sorted along the line of the sight beginning with the farthest one, which involves huge computational costs. In the case of the Borobudur temple, it is impractical to achieve interactive visualization with depth sorting, so the stochastic point-based rendering mechanism is used as a stochastic algorithm without depth sorting for transparent visualization. This work achieves a fused transparent visualization of the southeast corner of the Borobudur temple and the stochastic point-based rendering mechanism provides a promising result with a correct depth feel of 10^{10} points in a few seconds.

2. RELATED WORK

3D reconstruction of cultural heritage has gained a lot of attention in recent years. With laser-scanned data of intact cultural heritage objects, it is now possible to efficiently and flexibly obtain or reconstruct digital data of cultural heritage (Nuttens et al., 2011). Besides, as raw scanning data contains noise in the boundary, there are many manual modeling methods with various 3D reconstruction tools (Park et al., 2014). For many defective cultural heritage objects which are partially damaged, point generation approaches are widely applied (Hermoza, Sipiran, 2017, Lu et al., 2011). However, many cultural heritage objects are no longer available for acquiring 3D information due to irreversible damage. In this situation, image-based 3D reconstruction is needed and there are many methods that use multiple images to reconstruct 3D models (Kersten, Lindstaedt, 2013, Ioannides et al., 2013, Kyriakaki et al., 2014). However, only a single monocular photo per object remains prevalent in many cases. In this case, manual modeling methods are impractical for large-scale cultural heritage sites such as Borobudur, which is why an efficient reconstruction method from a single image is required.

The proposed method based on monocular depth estimation is used to map intensity into depth value from a single image, which is an ill-posed problem for its inherent ambiguity. Hand-crafted features and probabilistic graphical models are mainly used to tackle the monocular depth estimation problems in classical methods (Saxena et al., 2005). Recently, many studies using deep learning have achieved remarkable advances in depth estimation tasks. Deep learning is an efficient approach especially suitable for the relief reconstruction task carried out in this study. The common feature inside the reliefs can be learned and extracted by neural networks in order to eliminate manual workload. Most of the related studies involved working on indoor or outdoor scenes in which the depth was expressed in meters (Eigen et al., 2014, Laina et al., 2016). In one of the works (Pan et al., 2018), such depth prediction network was applied to relief-type data and its estimation possibility was proved in centimeters. However, that network structure contained two fully connected layers, in which a great number of parameters limited the resolution of the output depth map. In this paper, a fully convolutional network is used to increase the resolution of the output and reduce computational costs. Since over-fitting might occur during training owing to the complexity and the limited quantity of the data set as reported in (Pan et al., 2018). To avoid this issue, the proposed method also applies a residual structure inspired by the residual network ResNet.

Many studies propose a visualization method of point clouds based on opaque rendering (Kersten, Lindstaedt, 2013). However, in the case of transparent fused visualization of multiple point clouds, such studies are not appropriate. For the transparent visualization method for large-scale point clouds, the pioneering approach of (Zwicker et al., 2002) suffers from a large computational cost due to the depth-sorting process involved. On the other hand, the transparent fused visualization of large-scale point clouds has been barely studied. The stochastic point-based rendering used in this paper enables the rapid, precise, and interactive transparent rendering of large-scale point clouds (Tanaka et al., 2016). It achieves an accurate depth feel by employing a stochastic algorithm without depth sorting. In this study, the stochastic point-based rendering is applied to both the photogrammetry scanning point cloud and the reconstructed point cloud in order to achieve fused transparent visualization.

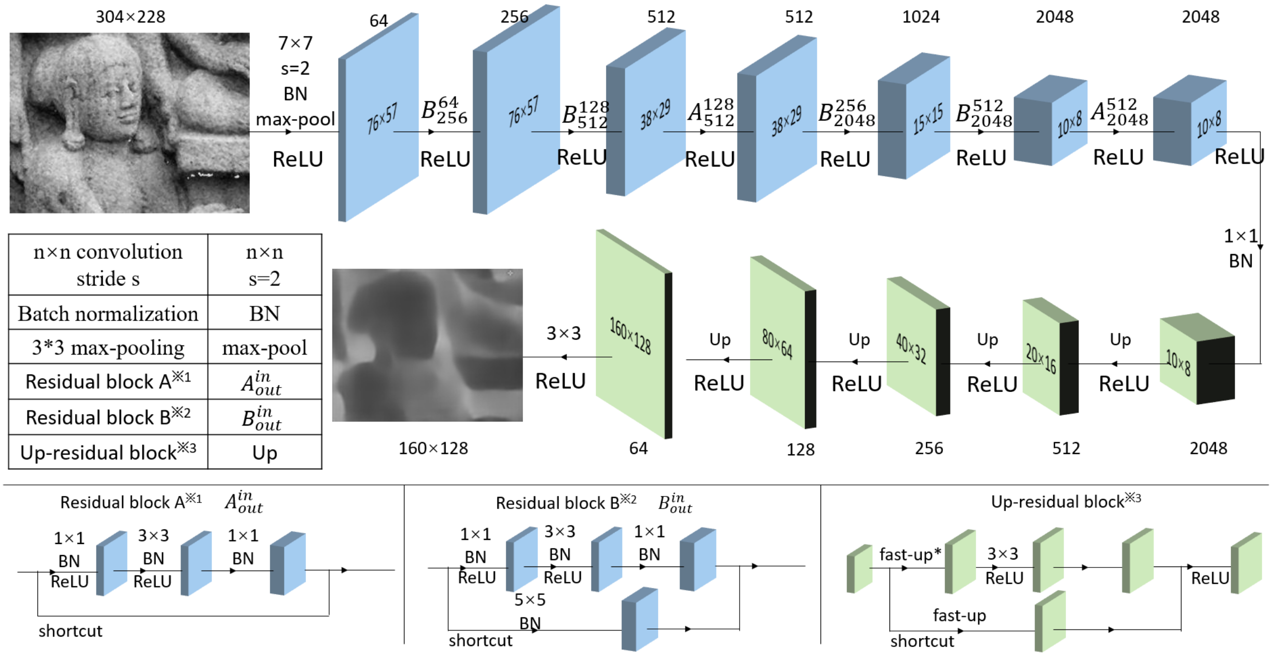


Figure 2. Network architecture of the depth estimation model. The proposed architecture applies ResNet-50 for down-sampling convolutional layers (blue parts). The fully connected layers are replaced by up-residual blocks (green parts) as described in (Laina et al., 2016). The fast-up in up-residual blocks is explained in Fig. 3.

3. METHOD

This section first describes the proposed 3D reconstruction method based on a depth estimation network, providing a detailed explanation of the network structure and the details of previous and post data processing. The transparent visualization method called the stochastic point-based rendering is then briefly introduced.

3.1 3D Reconstruction Method

To reconstruct the complete Borobudur temple into point cloud data, the 3D data of visible parts had to be merged with the 3D data of the buried Karmavibhangga reliefs. For the visible parts, photogrammetry scanning was used to obtain the 3D coordinator as well as color information. For the hidden parts, the only information that could be obtained is the gray-scale monocular photo containing 2D coordinator and intensity. Therefore the first step was to estimate the value of the axis *Z*, that is, the distance between the point and camera. Herein, a depth estimation neural network was used to map the intensity into a depth value. The photogrammetry scanning data was used as training data after separating its information into pairs of monocular photos and depth maps. The details of the depth estimation network will be provided in Section 3.2.

After the depth map was predicted from the single monocular photo, the value of axis *Z* could be obtained by a linear transformation from the depth value of each point. With the 3D coordinates and intensity, the gray-scale 3D points of the hidden Karmavibhangga reliefs were reconstructed from the old photos. The information about the camera or photography environment of the old photos is unavailable for reconstructing the 3D points under perspective projection rules. Moreover, the depth of the relief is much smaller than the photographic distance where its photo was taken and the relief is rectangular and the figures on it are extended only slightly from the background as shown in Fig. 1. Thus, we approximate that the old photo is a parallel projection of the hidden relief.

3.2 Depth Estimation Network

As convolutional neural networks decrease the resolution of input images during progressive convolutions and pooling strategies, an up-sampling strategy is required for high-resolution output in tasks such as depth estimation. The study in (Pan et al., 2018) uses a fully connected layer as the last layer and reshape the output into 55 × 74 resolution. The fully connected layers limit the output resolution for a large number of weights causing a large computational cost. Instead of fully connected layers, the present work uses deconvolutional layers with higher resolution output feature maps than their inputs.

The proposed architecture is shown in Fig. 2. The resolution of input monocular images with corresponding depth maps is 304 × 228. The resolution of the final output is 160 × 128, which is about half of the input resolution. The blue boxes on the top represent the feature map output by convolutional layers. This part uses the same architecture with ResNet-50 and is initialized with pre-trained weights. ResNet makes it possible to create deeper networks without facing vanishing gradients with a design of residual blocks. The residual blocks use a skip architecture with a shortcut for two convolutions and fuse the final outputs. The two residual blocks used in this architecture are shown at the bottom of Fig. 2.

The green parts at the bottom represent feature maps output by deconvolutional layers. The up-residual blocks have the same concept as the residual block B introduced in the blue parts. In the up-residual blocks, a fast-up convolution was used, as described by (Laina et al., 2016). The main idea of fast-up convolution is to avoid operations on zero pixels after conventional unpooling (Shi et al., 2016). In a feature map after unpooling, only certain locations are multiplied with non-zero values. As shown in Fig. 3, a 5 × 5 filter can be divided into four sets of weights based on the location(A,B,C,D). The sizes of four new filters are 3 × 3 for A, 3 × 2 for B, 2 × 3 for C and 2 × 2 for D based on the original filter. By interleaving the four final output feature maps by four sets of weights, the same output as the original output of conventional unpooling can be achieved. Compared to using a 5 × 5 filter on an 8 × 8 unpooled feature map, the fast-up convolution operation provides a more efficient way to use four sets of smaller filters on the original 5 × 5 feature map before unpooling by avoiding all zero multiplication.

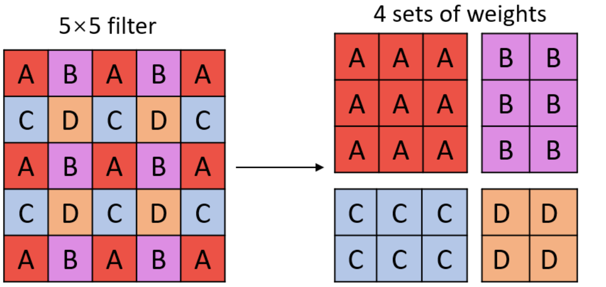


Figure 3. Fast-up convolution concept.

Only one 3×3 max-pooling layer is used at the beginning of the network to decrease input resolution. Besides, at the end of both blue and green parts, dropout layers are added to avoid over-fitting. After every convolution, batch normalization as ResNet is included to make further deeper networks. The work of (Pan et al., 2018) suffered from over-fitting using as a result of using an architecture of only eight hidden layers. Residual blocks and batch normalization make a great contribution to deepening the architecture on the limited data set used in this study.

The reverse Huber called BerHu (Lambert-Lacroix, Zwald, 2012) is used as the loss function for the proposed method in this work. As shown in Eq. 1, BerHu is a combination between the L_2 loss function (see Eq. 3) and L_1 (see Eq. 4) loss function. When $x < [c, c]$, BerHu is equal to L_1 loss and L_2 outside this range. Where there is a switch from L1 to L2, the function is continuous and first-order differentiable at point c. During training, if y^* represents predictions and y represents the ground truth, $B\left(y^* - y\right)$ is computed in every gradient descent step. Let i be index pixels over each image in the current batch, while c is set as shown in Eq. 2.

$$B\left(x\right) = \begin{cases} |x| & |x| \leq c \\ \frac{x^2+c^2}{2c} & |x| > c \end{cases} \tag{1}$$

$$c = \frac{1}{5} \max_i \left(|y_i^* - y_i|\right) \tag{2}$$

$$L_2\left(y^* - y\right) = \parallel y^* - y \parallel_2^2 \tag{3}$$

$$L_1(y^* - y) = |y^* - y| \tag{4}$$

L_2 loss is widely used in regression problems for minimizing the squared Euclidean norm. Due to the L_2 term, L_2 loss puts high weight to samples with a high residual. Meanwhile, L_1 has stronger influence on samples with smaller residual than L_2. Therefore, BerHu provides a good balance between L_1 and L_2 in the given task.

3.3 Transparent Visualization Method

This section introduces a transparent visualization method called the stochastic point-based rendering. Its details will be provided in Section.3.3.1, showing how a fused transparent visualization was achieved with the photogrammetry scanning point cloud of visible parts and the reconstructed point cloud of the hidden reliefs. Section.3.3.2 describes the fused visualization procedure for applying this method to multiple point clouds.

3.3.1 The Stochastic Point-based Rendering: The proposed stochastic point-based rendering is a transparent visualization method based on a stochastic algorithm without the need for depth sorting.

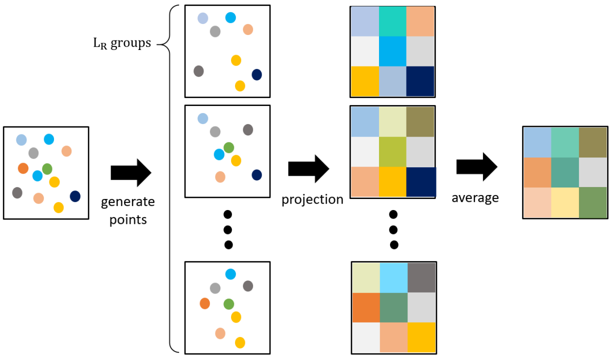


Figure 4. The stochastic point-based rendering.

The procedure of the stochastic point-based rendering method can be divided into three steps as shown in Fig. 4:

- **Step 1:** Prepare multiple subgroups of point clouds from original point clouds, each of which describes an equivalent but statistically independent surface. The point density of each group should be the same. The number of subgroups is denoted as L.
- **Step 2:** For each group in STEP 1, project its constituent 3D points onto the image plane to create an intermediate image. In the projection process, the point occlusion is considered per pixel. A total of L intermediate images are obtained.
- **Step 3:** Average the L intermediate images created in STEP 2 to make the final transparent image.

According to the above steps, in this method, parameter L controls image quality because it represents the number of averaged intermediate images. Besides, the opacity control of point clouds is not based on any depth-sorting procedures. Consider that s is the point sectional area whose image overlaps only one pixel and S is the area of the local surface segment that contains n points in total, then surface opacity in each local surface segment takes the following value:

$$\alpha = 1 - \left(1 - \frac{s}{S}\right)^n \tag{5}$$

According to Eq. 5, by tuning the local number of points, n can control local surface opacity α. By applying point-number adjustment, n, the chosen α can be realized. The set of points is uniformly eliminated to a small one in case of an unexpectedly large n. On the other hand, in the case of a small n, the number of points can be increased by creating a proper number of copies of randomly selected points so there is no need to add new points to the raw point cloud data. In this case, if parameter L is set to be large enough, each copy will approximately belong to a different subgroup created in STEP 2. Generally, parameter L is set to a few hundred of subgroups.

3.3.2 Fused Visualization of Multiple Point Clouds: In the case of fused visualization of multiple point clouds, the proposed stochastic point-based rendering provides a straightforward solution. After the point-number adjustment is applied to each point cloud, visualization can be achieved by simply merging the results.

- **Step 1:** Before executing the four steps of the stochastic point-based rendering, choose a user-defined value of opacity α for each point cloud.
- **Step 2:** The point-number adjustment procedure is applied to realize the chosen α.
- **Step 3:** By merging the adjusted point clouds, a unified point cloud can be created.

Therefore, to emphasize a specified point cloud, higher opacity values can be assigned before executing the stochastic point-based rendering procedure.

4. EXPERIMENTS

This section introduces the relief data set used to evaluate the method proposed in Section.4.1. The details of the implementation of the proposed method will be described in Section.4.2, whereas the quantitative and qualitative results obtained by the proposed model will be explained in Section.4.3.

4.1 Relief Data Set

As explained in Section.3, the input of the proposed network are pairs of monocular photos and depth maps. As the original data is point cloud data, 3D coordinates and color information had to be separated into a monocular photo and the corresponding depth map. The intensity in the depth map was calculated from the value of the axis Z by a linear transformation. Intensity was set in the range of 0 to 255. In a specified point cloud of a relief, if the maximum value of the axis Z is Z_{max} and the minimum value of the axis Z is Z_{min}, the intensity d in the depth map follows Eq. 6:

$$d = \frac{Z - Z_{min}}{Z_{max} - Z_{min}} \times 255 \tag{6}$$

Following the above equation, 11 pairs of monocular photos and depth maps were made. To train the deep estimation neural network, these large-resolution images were cut into 4,087 pairs of image patches. The following data augmentation methods were applied to the data set: rotation, flips, noise, and blurry. The final data set contained 44,957 pairs of monocular photos and corresponding depth maps which were 11 times the quantity of the origin.

4.2 Implementation Details

TensorFlow (Abadi et al., 2016) was used for the implementation of the proposed network. The model was trained on a single NVIDIA GeForce GTX 1080Ti with 12 GB of GPU memory. The weights of the blue parts in the proposed network were initialized by ResNet pre-trained on the image classification data set. Besides, the weights of the green parts were randomly initialized. The probability of dropout layers added at the end of both blue and green parts was set to 0.5. The model was trained using an Adam optimizer with a learning rate of 0.001. The model was trained on the relief data set with a batch size of four for approximately 100 epochs.

4.3 Architecture Evaluation

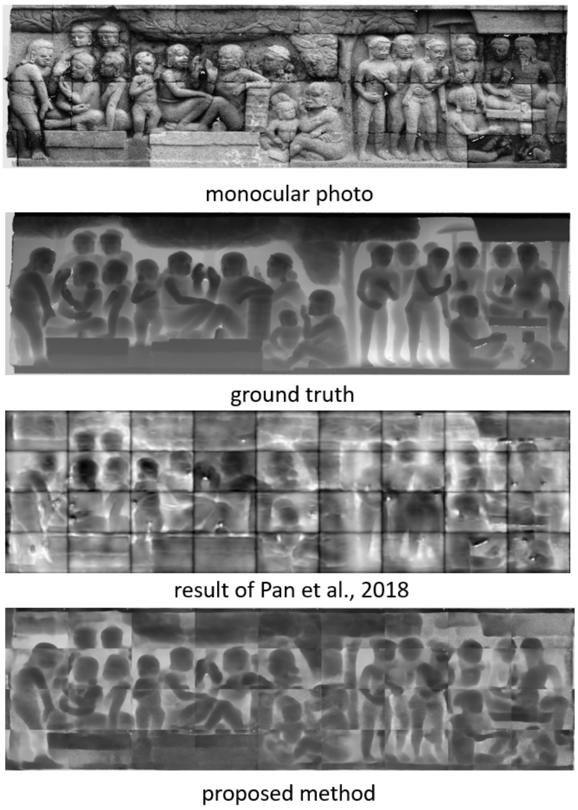
As described in Section.1, the reliefs which need reconstruction belong to a section called Karmavibhangga. Out of the 160 Karmavibhangga reliefs, there are four reliefs on the southeast temple wall that have remained visible and the other 156 reliefs were buried under the stones. To evaluate the proposed method, one relief from the visible four panels of the Karmavibhangga was chosen for the quantitative and qualitative experiments (see Fig. 5). Although the parts that need reconstruction are the hidden reliefs buried by stones, it was impossible to measure their accuracy directly due to the unavailable ground truth. It is a concession but also a sensible choice to evaluate the proposed model on the visible parts instead of the hidden parts.

The results of the quantitative analysis are shown in Tab. 1. 2D and 3D error metrics were applied to measure the accuracy of the proposed method. First, as the reconstructed point clouds are based on the depth prediction, to evaluate the accuracy of the predicted depth map, 2D error metrics were applied. The ratio of pixels correctly labeled in predicted depth map is calculated by $\alpha1$, $\alpha2$ and $\alpha3$ (Eigen et al., 2014). As the depth value is always positive or zero (in the range of 0 to 255) with highly skewed distribution which makes the symmetric loss function such as RMSE not applicable enough, a logarithmic transformation (RMSElog) is applied to obtain a less skewed distribution. Second, to evaluate the real distance (into meters) the study also measured the 3D distance between the reconstructed point clouds and the ground truth. The 3D cloud-to-cloud distance between each corresponding point was calculated and visualized in a heat map. The cloud-to-cloud distance represents the real error because the point clouds were reconstructed into the real size of Borobudur reliefs. For comparison experiments, these error metrics were applied to the proposed method and the study of (Pan et al., 2018).

As shown in Tab. 1, we here compare the result of the proposed method (Exp2) to the previous work (Exp1) which is proposed in (Pan et al., 2018). For the 2D predicted depth maps, the proposed method based on ResNet has been substantially improved for each error metric applied. The relief which was chosen as

Error metrics			Exp1	Exp2
2D	Higher is better	$\alpha1 < 1.25$	0.25047	0.47393
		$\alpha2 < 1.25^2$	0.45433	0.77222
		$\alpha3 < 1.25^3$	0.60945	0.88119
	Lower is better	RMSE	10.24803	10.07052
		RMSElog	0.41194	0.25000
3D	c2c distance (in meters)		0.01498m	0.00890m

Table 1. Quantitative comparison results.


monocular photo

ground truth

result of Pan et al., 2018

proposed method

Figure 5. Depth estimation result.

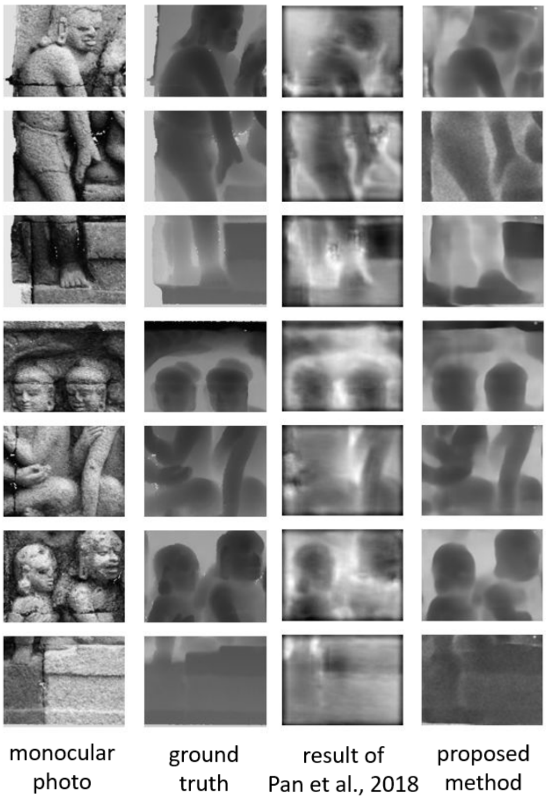


Figure 6. Patch-wise depth estimation result.

test data is shown in Fig. 5, along with the ground truth and results of comparative experiments. As the input of the neural network is fixed to 304×228, the test image is cut into patches for depth estimation. The most meaningful results are also presented in Fig. 6. The study in (Pan et al., 2018) uses multi-scale architecture to combine global and local information. While a shallow sub-network provides sharper boundaries, a lot of noise also appears in the final output. As shown in the depth map, the intensity in the output in (Pan et al., 2018) is unstable. In the case of the proposed method, the deeper architecture and residual operation provide a much better prediction result. The value of the intensity is stable, which means that the value of the intensity within a small area is approximately equal. In the background of the relief, when the method of (Pan et al., 2018) fails as a lot of areas turn to intense bright, the result of the proposed method remains smooth and correct.

The reconstructed point cloud provides a 3D understanding with the correct depth feel of the relief, as shown on the top of Fig. 7. Besides, the distances of each of its points relative to the ground truth were computed in meters. The mean distance obtained by the proposed method between the reconstructed point cloud and the ground truth point cloud was approximately 0.008 m, as shown in Tab. 1. Compared to 0.015 m provided in the work of (Pan et al., 2018), it is approximately a one-half reduction. As the real Borobudur relief is 2.7m wide, 0.92m high and 15cm in depth, the accuracy of the proposed method is approximately 95%. The error distance of each point is visualized in the heat map shown in the middle of Fig. 7. As the color turns from blue to red, the error distance increases. The accuracy of human figures is the highest, as they are mostly covered with blue. Moreover, the error distance of almost all points is lower than 0.01 m, while only a few points reached an error distance greater than 0.02 m.

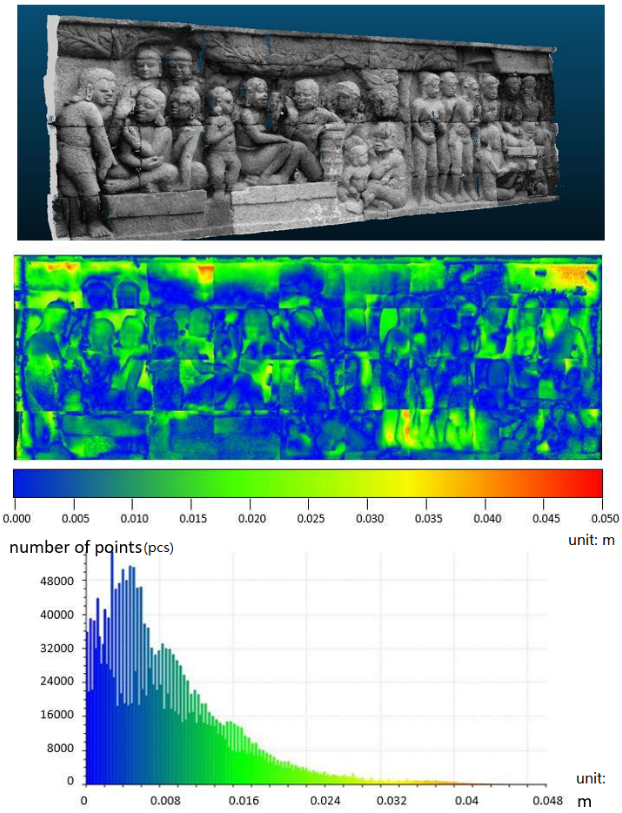


Figure 7. Reconstruction result (test data).

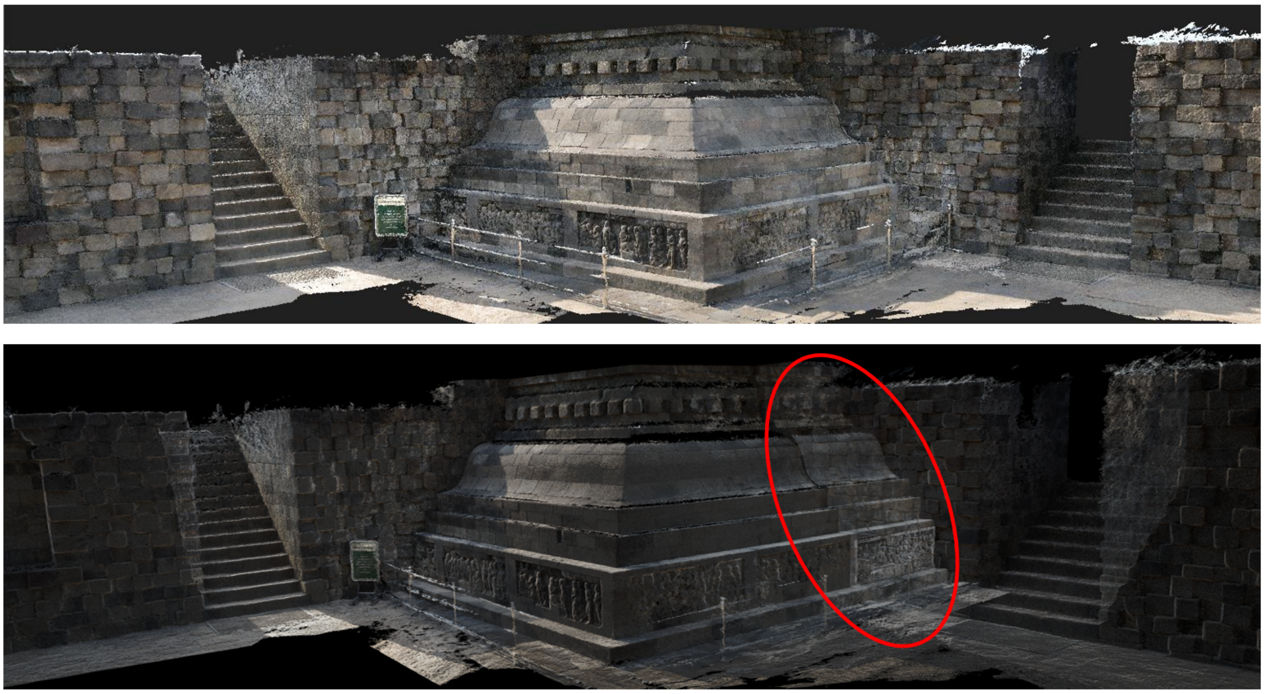


Figure 8. The fused transparent visualization result of the southeast temple wall in Borobudur. The opaque rendering result is shown on the top while the fused transparent visualization result is shown on the bottom. The reconstructed part is inside the red circle. The high resolution see-through image of this part is shown in Fig. 10.

5. RESULTS

This section presents the reconstruction point cloud of the hidden reliefs. Note that it is not possible to perform any quantitative analysis on the reconstructed model due to the unavailability of the ground truth. As the point cloud is reconstructed from a single gray-scale monocular photo (see Fig. 1), the number of points is equal to image resolution. Fig. 9 shows a reconstructed relief in Karmavibhangga indexed by number 18. This relief is buried under the stone wall which just next to one side of the southeast corner in the Borobudur temple. The reconstructed point cloud contains 1,663,488 points and the color of each point follows the intensity in the old monocular photo. The relief was reconstructed using the trained network within a few seconds and the surroundings were manually added to approach its original appearance.

After the reconstruction, the reconstructed data were combined with the photogrammetry scanning data of the southeast temple corner by merging them into the same coordinates. Then the stochastic point-based rendering was applied to both point clouds. Based on different density of the point clouds, the opacity α was set to 0.8 for the photogrammetry scanning data and 0.2 for the reconstructed point cloud. The fused transparent visualization result is shown in Fig. 8. With the proposed method, it is possible to see through the stone wall and figure out the 3D appearance of the hidden relief. The proposed method provides high-quality see-through imaging as shown in Fig. 10. In the case of large-scale cultural heritage, the stochastic point-based rendering provides transparent visualization at an interactive speed, which is very important for further application in virtual reality or in walk-through displays.

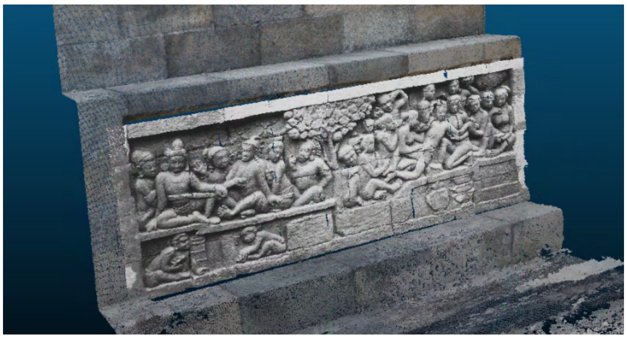


Figure 9. Reconstruction result from the old photo.

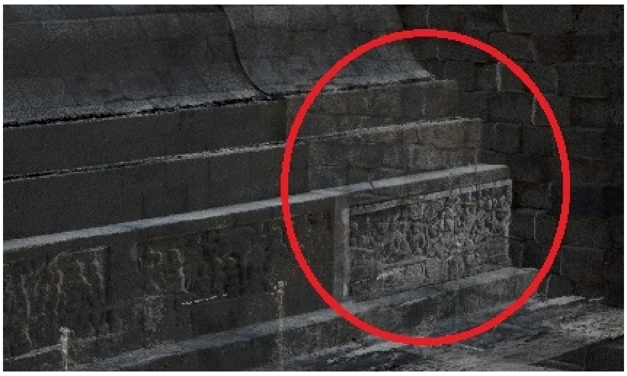


Figure 10. Fused transparent visualization result (zoomed in).

6. CONCLUSIONS

This work presented an efficient method for fused transparent visualization of incomplete cultural heritage based on a monocular depth estimation neural network and stochastic point-based rendering. Given a monocular photo, the proposed method can efficiently reconstruct 3D surface points and provide transparent visualization with proper depth feel in a few seconds. The proposed depth estimation neural network is a single-scale fully convolutional neural network that follows the ResNet-50. This architecture provides a deeper network for further feature extraction and a larger resolution of the output.

For the purposes of the evaluation of the method and its applicability, the proposed method was applied to the Borobudur temple. The hidden Karmavibhangga reliefs were reconstructed into point clouds and a fused transparent visualization was achieved with the photogrammetry scanning point cloud data of the southeast temple corner and the reconstructed data of the hidden Karmavibhangga reliefs. As a result of the quantitative and qualitative experiments, the accuracy of the reconstructed point clouds is 95%, which is an increase of a 5% increase of the study in (Pan et al., 2018). The fused transparent visualization provides a 3D understanding with the correct depth feel of the original appearance of the Borobudur temple before its restoration in 1890.

ACKNOWLEDGEMENTS

In this paper, the images of the Borobudur temple are presented and with the permission of the Borobudur Conservation Office and Research Center for Area Studies (P2W) of the Indonesian Institute of Sciences (LIPI). Additionally, photogrammetry scanning point cloud data was provided by the Nara National Research Institute for Cultural Properties. We deeply thank these three institutions for their generous cooperation. This work is partially supported by JSPS KAKENHI Grant Number 19KK0256 and the Program for Asia-Japan Research Development (Ritsumeikan University, Japan).

REFERENCES

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mane, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viegas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. http://arxiv.org/abs/1207.6868.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, 2366–2374.

Hermoza, R., Sipiran, I., 2017. 3D Reconstruction of Incomplete Archaeological Objects Using a Generative Adversarial Network. http://arxiv.org/abs/1711.06363.

Ioannides, M., Hadjiprocopi, A., Doulamis, N., Doulamis, A., Protopapadakis, E., Makantasis, K., Santos, P., Fellner, D., Stork, A., Balet, O., Julien, M., Weinlinger, G., Johnson, P. S., Klein, M., Fritsch, D., 2013. Online 4d reconstruction using multi-images available under open access. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, 2(5/W1), 169–174.

Kersten, T. P., Lindstaedt, M., 2013. Automatic 3D Object Reconstruction from Multiple Images for Architectural, Cultural Heritage and Archaeological Applications Using Open-Source Software and Web Services Automatische 3D-Objektrekonstruktion aus digitalen Bilddaten für Anwendungen in Archit. *Photogrammetrie - Fernerkundung - Geoinformation*, 2012(6), 727–740.

Kyriakaki, G., Doulamis, A., Doulamis, N., Ioannides, M., Makantasis, K., Protopapadakis, E., Hadjiprocopis, A., Wenzel, K., Fritsch, D., Klein, M., Weinlinger, G., 2014. 4D Reconstruction of Tangible Cultural Heritage Objects from Web-Retrieved Images. *International Journal of Heritage in the Digital Era*, 3(2), 431–451. http://journals.sagepub.com/doi/10.1260/2047-4970.3.2.431.

Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N., 2016. Deeper depth prediction with fully convolutional residual networks. *Proceedings - 2016 4th International Conference on 3D Vision, 3DV 2016*, 239–248.

Lambert-Lacroix, S., Zwald, L., 2012. The BerHu penalty and the grouped effect. http://ljk.imag.fr/membres/Laurent.Zwald.

Lu, M., Zheng, B., Takamatsu, J., Nishino, K., Ikeuchi, K., 2011. Preserving the Khmer smile: classifying and restoring the faces of Bayon. *Proceedings of the 12th International conference on Virtual Reality, Archaeology and Cultural Heritage*, 161–168. https://dl.acm.org/citation.cfm?id=2384521.

Nuttens, T., De Maeyer, P., De Wulf, A., Goossens, R., Stal, C., 2011. TS06E-Laser Scanning and Photogrammetry-5267 Terrestrial Laser Scanning and Digital Photogrammetry for Cultural Heritage: an Accuracy Assessment Terrestrial Laser Scanning and Digital Photogrammetry for Cultural Heritage: an Accuracy Assessment. Technical report.

Pan, J., Li, L., Yamaguchi, H., Hasegawa, K., Thufail, F. I., Tanaka, S., 2018. 3D Reconstruction and Transparent Visualization of Indonesian Cultural Heritage from a Single Image. *Proc. The 16th EUROGRAPHICS Workshop on Graphics and Cultural Heritage, Vienna, November 12-15, 2018.*

Park, J. H., Muhammad, T., Jae-Hong, A., 2014. The 3D reconstruction and visualization of Seokguram Grotto World Heritage Site. *Proceedings of the 2014 International Conference on Virtual Systems and Multimedia, VSMM 2014*, 180–183.

Saxena, A., Chung, S. H., Ng, A. Y., 2005. Learning depth from single monocular images. *Advances in Neural Information Processing Systems*, 1161–1168.

Shi, W., Caballero, J., Huszar, F., Totz, J., Aitken, A. P., Bishop, R., Rueckert, D., Wang, Z., 2016. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem, 1874–1883.

Tanaka, S., Hasegawa, K., Okamoto, N., Umegaki, R., Wang, S., Uemura, M., Okamoto, A., Koyamada, K., 2016. See-through Imaging of Laser-scanned 3D Cultural Heritage Objects based on Stochastic Rendering of Large-scale Point Clouds. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.*, III-5, 73–80.

Zwicker, M., Pfister, H., Van Baar, J., Gross, M., 2002. EWA splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3), 223–238.