Jiao Pan. 2024. Chapter 4: Fused 3D Transparent Visualization for Large-Scale Cultural Heritage Using Deep Learning-Based Monocular Reconstruction. In Satoshi Tanaka (ed.), *Digital Archiving of Cultural Properties Based on Advanced ICT and Utilization of the Archived Data*. Osaka: Asia-Japan Research Institute, Ritsumeikan University. ISBN 978-4-910550-26-8

Chapter 4
Fused 3D Transparent Visualization for Large-Scale Cultural Heritage Using Deep Learning-Based Monocular Reconstruction

Jiao PAN

Figure 1. Borobudur Temple
Source: Author

The visualization of digital archives is increasingly important in the preservation and also the analysis of cultural heritages. With 3D scanning technology, it is possible to efficiently acquire and preserve digital data of existent cultural heritages. For some cultural heritages that no longer exist, multiple image-based methods can also be used to reconstruct 3D models.

However, there are many cases where only one monocular photo remains for one cultural heritage. In this case, as 3D scanning and multiple image-based methods cannot be applied, the ability to reconstruct a 3D model from a single image is urgently required.

Figure 1 shows the Borobudur Temple, a UNESCO World Heritage site Feener introduced in Chapter 1. It is the largest Buddhist temple in the world. On the temple wall, there is the most complete collection of Buddhist reliefs in the world, containing more 1,400 panels of reliefs.

Each relief tells a different story, accurately illustrated by many objects. Based on different themes, these reliefs can be divided into five sections. One section is named the Karmavibangga, the Hidden Foot. Due to restoration work 200 years ago, almost the entire section of 160 panels of the Karmavibangga is buried under stones and is no longer visible to visitors except for four panels on the southeastern temple wall.

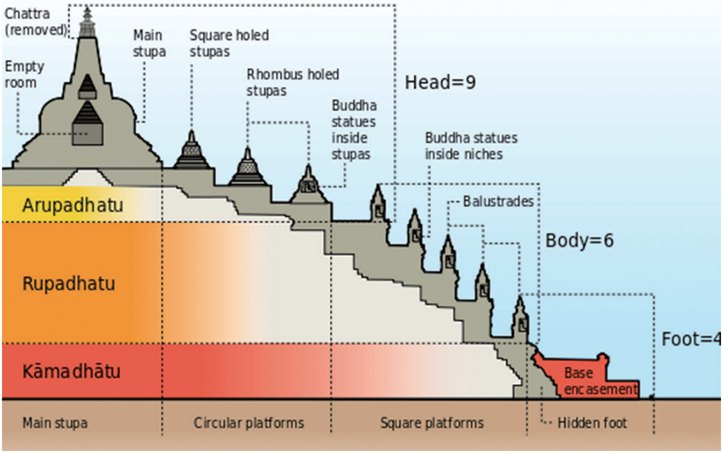


Figure 2. Section of Borobudur showing the hidden foot (in red)
Source: Autho

In Figure 2, the red box represents the southeast corner of the temple. Figure 3 is a photo of the corner, and you can see that only four remaining reliefs on the temple wall are visible, and all the other reliefs on the first level, which are shown in yellow, are covered by stones.

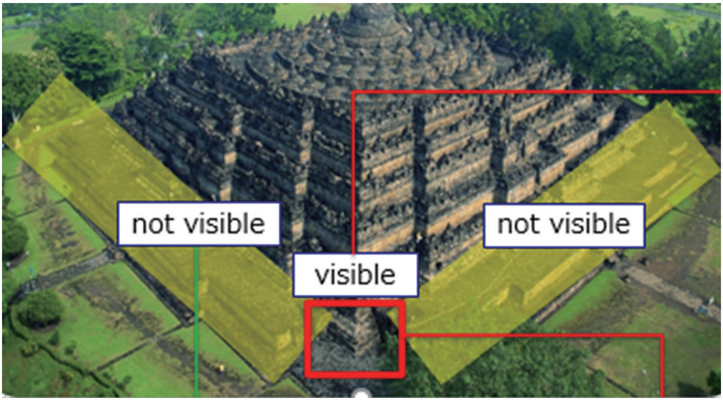


Figure 3. Exposed corner of the temple
Source: Author

To record the visible parts, we can simply use 3D photographic scanning and visualize the scanning data. However, we need to find another way to get the image of the buried parts, as only some grayscale photos taken in 1890 remain. Figure 4 shows one of these grayscale monocular photos and there is only one photo remaining for each panel.

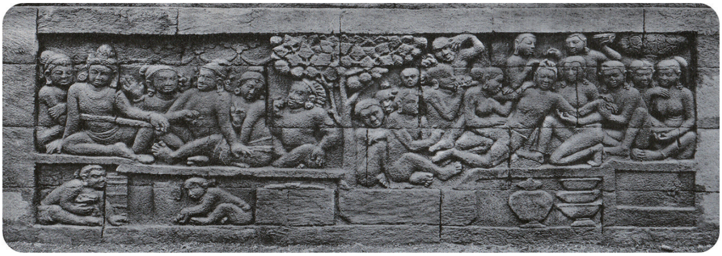


Figure 4. 1890 photo of the buried relief
Source: Author

As these photos are the only remaining record, conventional digitizing by 3D laser or photo scanning multiple image-based methods cannot be applied in this case. So, we proposed using monocular depth estimation based on deep learning to reconstruct a 3D model directly from the old monocular photo. Fortunately, all the old photos taken in 1890 are still available.

Visualization Method

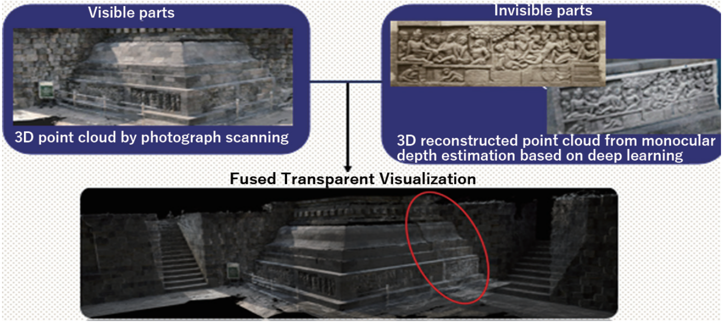

Visible parts
3D point cloud by photograph scanning
Invisible parts
3D reconstructed point cloud from monocular depth estimation based on deep learning
Fused Transparent Visualization

Figure 5. Flowchart of proposed visualization method
Source: Author

Figure 5 is a flowchart of our proposed method. For the visible parts, we use photograph scanning to obtain 3D point clouds. For the buried parts, which are not visible, we use deep learning to estimate the depth from the monocular photo and reconstruct it into point clouds. Then, we can fuse these two kinds of point clouds together and provide a transparent visualization for the entire Borobudur Temple.

This method is efficient and accurate for providing actual 3D visualization directly from a single monochrome photo of the image, especially for relief-type cultural heritages.

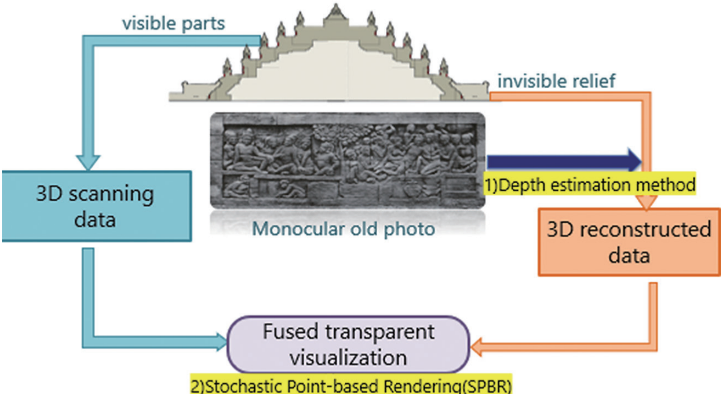

visible parts
invisible relief
3D scanning data
Monocular old photo
1)Depth estimation method
3D reconstructed data
Fused transparent visualization
2)Stochastic Point-based Rendering(SPBR)

Figure 6. Fused 3D transparent visualization for Borobudur
Source: Author

Figure 6 shows an overview of the method we employed. To achieve this fused 3D transparent visualization, we mainly use two methods. The first one is the depth estimation method which is used to reconstruct the point clouds of the buried parts. The second one is the transparent visualization method, the Stochastic Point-based rendering (SPBR) that our lab developed.

When there is no extra information to use, we need to estimate depth by using deep learning to reconstruct the 3D model from a monocular 2D image. With this depth map, we can reconstruct the 3D shape of the reliefs because the value of each pixel in the depth map represents the depth distance between the point and the camera.

Deep Learning Method

Next, I will briefly introduce how deep learning works in our case by explaining how we get the estimated depth map from the monograph photo. It is a machine-learning method based on deep neural networks, which has become more and more popular in recent years. In computer science, it is named neural network because its structure is similar to the neurons in the human brain. Figure 7 shows a drawing of a basic neural network. It has input layers, which is what you have, and output layers, which is what you want to obtain.

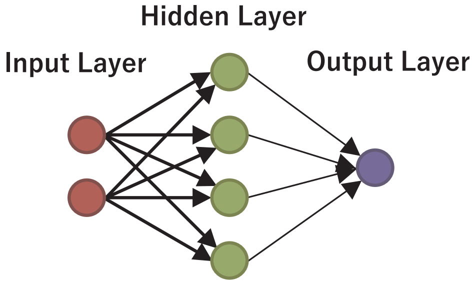


Figure 7. A simple artificial neural network
Source: Author

Moreover, the hidden layers are usually designed to be very deep and complicated, so that we can learn and solve difficult problems with them. In Figure 7, the hidden neural has only one layer, but it usually has hundreds or even thousands of layers. We use the monocular images as the input and the depth map as the output. In order to train the model, we need to tell the model direct answers to the questions; that is, we need to use ground truth of the old photos during the training process. In this step, we need a great number of relief pictures and the ground truth, which are the depth maps for the training set. First, we use photographic scanning to scan the visible parts and obtain our original data set, which is point cloud data of those visible four reliefs in Karmawibhangga. Next, we separate the information into pairs of images and depth maps, which is the label for our training. As the reliefs are large-scale, we cut the large images into patches and apply this augmentation method to them. While obtaining the individual scanning data, it is also necessary to reduce the effects of light conditions and noises, otherwise it will affect the results. Before our data is fed to the depth estimation network, median filter and batch normalization are applied to the data set to reduce the effect of different light conditions.

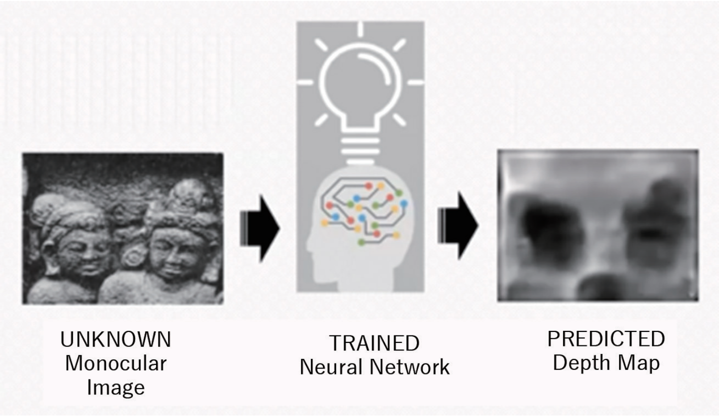
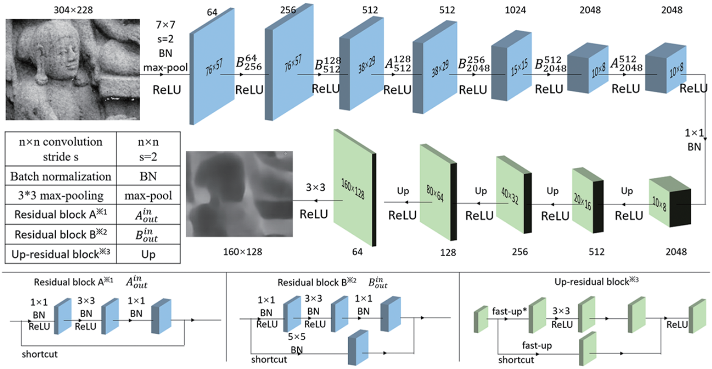


Figure 8. Completing the training process
Source: Author

When we have finished the training process, we can input the old photo of the hidden reliefs, which are totally unknown to the neural network. The training model will estimate the depth loss, and we can get an output depth map as in Figure 8.

Now, let me summarize the depth estimation method. To reconstruct the 3D model from a monocular photo, the most important difference between a 2D monocular photo and a 3D Point Cloud is that we do not have depth estimation in the photos, which is the value of the Z axis. In the first step, we apply a depth prediction neural network to map intensity to depth value. The network will provide us with a depth map where the value of each pixel contains the distance between the point and the camera. Once we obtain the depth map, point cloud data can be reconstructed by a linear transformation between the depth value and the Z axis. In Step 1, we need the relief data set to train the model because we apply a neural network. In our case, we use pairs of monocular photos and depth maps. We calculate the dataset from the scanning 3D points of the visible reliefs in Borobudur.

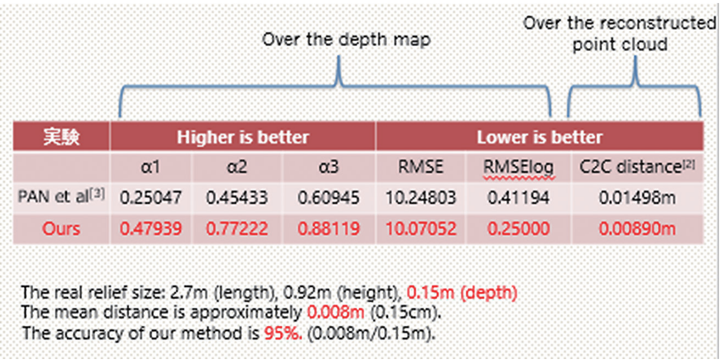


n×n convolution stride s	n×n s=2
Batch normalization	BN
3*3 max-pooling	max-pool
Residual block A※1	A_{out}^{in}
Residual block B※2	B_{out}^{in}
Up-residual block※3	Up

Figure 9. ResNet based depth estimation network
Source: Author

In our depth estimation network, we use more than 60 hidden layers based on a famous network named the ResNet. The blue blocks in Figure 9 are the encoders, which reduce the image resolution and extract the features, and they are based on ResNet. The green blocks represent the decoder. So, instead of the fully connected layers, we use the deconvolutional layers to recover the image resolution, and therefore, the outputs can be nearly half of the input. With this model, we can get a clear output depth map of the old photo.

For the evaluation, we made a quantitative comparison with our previous work. Figure 10 shows the qualitative results. The last column is the 3D error calculators over the reconstructed point cloud and the other is calculated over the depth maps. Significantly, the actual relief size is 0.15m, and our depth result is 0.089 centimeters. So, as we can see in the table, the accuracy of the proposal method is about 95%.

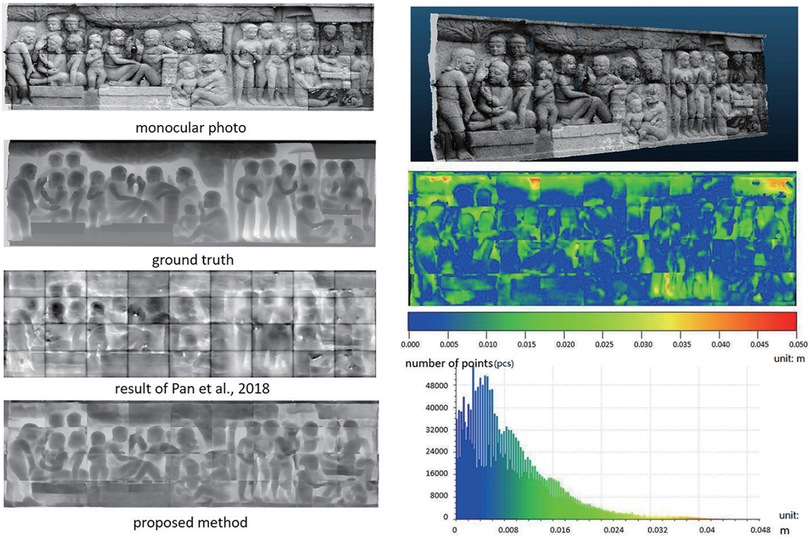

Over the depth map
Over the reconstructed point cloud

実験	Higher is better			Lower is better		
	α1	α2	α3	RMSE	RMSElog	C2C distance[2]
PAN et al[3]	0.25047	0.45433	0.60945	10.24803	0.41194	0.01498m
Ours	0.47939	0.77222	0.88119	10.07052	0.25000	0.00890m

The real relief size: 2.7m (length), 0.92m (height), 0.15m (depth)
The mean distance is approximately 0.008m (0.15cm).
The accuracy of our method is 95%. (0.008m/0.15m).

Figure 10. Quantitive Comparison Experiment
Source: Author

Figures 11 and 12 show the qualitative conversion results. Figure 10 shows the results of depth maps. The first row is the monocular photo, and the second is the ground truth. The last row is the result of the proposed method.



Figures 11and 12. Qualitative Comparison results
Source: Author

We can see the improvement here. On the right is the heatmap of the difference, which is calculated by C2C distance between the reconstructed data and the scanning data. As the color turns from blue to red, the error distance increases. Moreover, as the histography shows the error distance of almost all the points is lower than two centimeters.

After the reconstruction, we also applied a transparent visualization method known as Stochastic Point-based Rendering (SPBR) which our lab designs. It is a high-quality, see-through imaging mechanism for point cloud data. We have successfully applied it to many kinds of point cloud data, including cultural heritage. This method can achieve a transparent visualization result without any in-depth sorting, which means it does not suffer from large computation costs. This means that it is very suitable in our case because the scale of the point cloud of the entire Borobudur temple is exceptionally large.

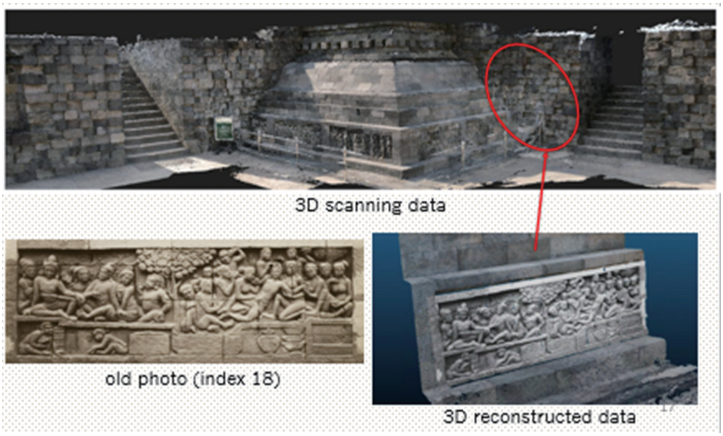


Figure 13. Fused Visualization Results
Source: Author

Figure 13 shows the visualization results by using SPBR. Here we use the 3D scanning data of the southeast temple corner and the reconstructed data of the old photo indexing number 18 as an example. The location of photo number 18 is behind the wall in the red ring and right next to the visible reliefs. The top picture shows the opaque visualization, and you cannot tell where the hidden relief is, but in the bottom picture we can look into the stones and figure out what the hidden relief looks like. When we zoom in you can see the details of the reliefs clearly.

Conclusion

In conclusion, this work provides an efficient method for fused transparent visualization of incomplete cultural heritages based on a monocular depth estimation neural network and stochastic point-based rendering. We applied our method to the Borobudur temple and visualized a corner as an example. The reconstruction accuracy achieved was 95% and the fused visualization and the field's transparent visualization provided us with promising results.

For future work, first we want to improve the reconstruction methods as we want to work on a larger dataset. We also want to design and use other models which are more suitable for the relief type data. Also, to improve the visualization results we are planning to apply edge-highlighting to the reliefs. Finally, we hope to complete the entire temple visualization.

References

Pan, Jiao, Weite Li, Liang Li, Kyoko Hasegawa, and Satoshi Tanaka. 2022. Deep Learning in Cultural Heritage: Improving the Visualization Quality of 3D Digital Archives. *Journal of the Asia-Japan Research Institute of Ritsumeikan University*, 4, 175–190.

Pan, Jiao, Liang Li, Hiroshi Yamaguchi, Kyoko Hasegawa, Fadjar I. Thufail, Brahmantara, and Satoshi Tanaka. 2020. Fused 3D Transparent Visualization for Large-scale Cultural Heritage Using Deep Learning-based Monocular Reconstruction. *ISPRS Ann. Photogramm. Remote Sens. Spatial Inf. Sci.(Proc. XXIV ISPRS Congress)*, V-2-2020, 989–996.

———. 2021. Integrated High-Definition Visualization of Digital Archives for Borobudur Temple. *Remote Sensing*, 13(24), 5024.

———. 2022. 3D Reconstruction of Borobudur Reliefs from 2D Monocular Photographs Based on Soft-edge Enhanced Deep Learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 183, 439–450.

Dr. Jiao PAN

Chapter 4. Fused 3D Transparent Visualization for Large-Scale Cultural Heritage Using Deep Learning-Based Monocular Reconstruction

Jiao Pan majors in Information Science and Engineering, with a keen focus on deep learning-based image processing and the 3D digitalization of historical artifacts. She earned her Ph.D. from the Graduate School of Information Science and Engineering at Ritsumeikan University in 2022. Following this, she held a position as a JSPS Postdoctoral Fellow until May 2023. Since July 2023, she has been serving as a lecturer at the University of Science and Technology Beijing. Among her notable publications are "3D Reconstruction of Borobudur Reliefs from 2D Monocular Photographs Based on Soft-Edge Enhanced Deep Learning," *ISPRS Journal of Photogrammetry and Remote Sensing* (January 2022, Issue 183, Pages 439–450), and "Integrated High-Definition Visualization of Digital Archives for Borobudur Temple," *Remote Sensing* (December 2021, Vol. 13, Issue 24, Page 5024).