EUROGRAPHICS Workshop on Graphics and Cultural Heritage (2023)
A. Bucciero, B. Fanini, H. Graf, S. Pescarin, and S. Rizvic (Editors)

Short Paper

AI Based Image Segmentation of Cultural Heritage Objects used for Multi-View Stereo 3D Reconstructions

H. Kutlu[1], F. Brucker[1], B. Kallendrusch[1], P. Santos[1], and D. W. Fellner[1,2]

[1]TU Darmstadt & Fraunhofer IGD, Germany
[2]Graz University of Technology, Institute of Computer Graphics and Knowledge Visualization, Austria

Abstract

Image segmentation (or masking) finds a very useful use case within 3D reconstruction of cultural heritage objects. The 3D reconstructions can be accelerated, reconstructing the object without any background noise. Conventional segmentation methods can calculate erroneous masks for certain objects and environments, which can lead to errors within the reconstruction: Parts of the 3D reconstruction may be missing or are incorrectly reconstructed, which contradicts adequate archiving. The automated iterative Multi-View Stereo (MVS) scanning process makes it necessary to obtain masks that reconstruct the object in the best possible way, regardless of the environment, the stabilizing mount, the color of the background and the object. In addition, it should not be necessary to tweak the best possible parameters for conventional masking procedures and to create masks manually. State-of-the-art artificial intelligence (AI) segmentation networks will be trained and applied to the MVS scans to verify the behavior of the associated 3D reconstructions and the automated iterative scanning process. In addition, a comparison between different AI segmentation networks and a comparison between conventional masking methods and AI segmentation networks is performed.

CCS Concepts

• ***Computing methodologies*** → ***Image segmentation; Reconstruction;*** • ***Hardware*** → *Scanners;*

1. Introduction and related work

The digitization of cultural heritage objects has become increasingly relevant in recent years. Natural disasters such as the recent earthquakes in Turkey, as well as wars, ensure that non-digitized cultural artifacts disappear without a way to archive them adequately. Additionally, the pandemic showed the importance of digitizing these objects for their respective museums and for access by the public. In addition, digitization makes it possible to conduct research with the artifacts without having to move them. However, 2D digital replicas cannot fulfill these requirements for all artifacts, requiring a strategy for a 3D digitization pipeline that is true to the geometry and material properties. In order to digitize the high number of these cultural objects in the best possible quality, the CultLab3D [SRT*14] was founded at the Fraunhofer Institute for Computer Graphics Research, which solves this challenge by developing an autonomous mass digitization pipeline. One of the challenges was the digitization of the geometries of arbitrary objects, which was solved by an automated iterative scanning approach using the CultArm3D [STD*20]. In this approach, autonomous images are taken from different positions of the object to reconstruct an intermediate 3D model using photogrammetry, which in turn is used to compute the next best possible camera positions. This is repeated until a certain quality can be expected in the final model. For the reconstruction of the intermediate 3D models, masks are used to speed up the reconstruction and at the same time avoid the occurrence of noise within the 3D models. If the masks are too inaccurate, the reconstructions calculates an incorrect 3D model. This leads to incorrectly calculated new camera positions for the acquisition of images, which can cause a scan abort. Image segmentation gets complex as soon as the objects match the background color. In addition, mounts are sometimes used during the scan to stabilize the objects, which can also be similar in color. Developing a general solution is not trivial, as it strongly depends on the respective object, since those are unique within the cultural heritage domain. Therefore, it was analyzed whether a generalizable solution can be found with the help of state-of-the-art AI segmentation networks.

Within the cultural heritage domain, image segmentation has already been performed on Chinese literati paintings to extract the fine and complex lines [ZZX*20]. These paintings are drawn on Xuan paper, where the lines are created by different mixing ratios of ink and water. Over time, these papers yellow, causing the fine lines to change, so classical clustering methods were no longer suitable for image segmentation. Multi-view fuzzy clustering is used with the assumption that the three RGB channels of an image can be considered as multi-view images. Automated segmentation was analyzed within cultural heritage imaging, where the challenge was to automatically remove people from images and estimate the missing information using a pre-trained model [MGC13]. Among other

DOI: 10.2312/gch.20231160

things, support vector machines were used to generate masks for the people to be removed.

The general concept of masking things out of images is not suitable for our case. Therefore, different AI-based solutions are evaluated and analyzed with the following contributions:

- Comparison of different AI networks for image segmentation.
- Comparison between conventional and AI-based segmentation.
- Application of the estimated masks within a scanning process to compare the 3D reconstructions.

2. Workflow and utilized networks

The automated iterative scanning pipeline, as shown in Figure 1, starts with a roughly covered image set of the object that is captured using a camera attached to a 6-degrees of freedom robot arm. Corresponding to a mass digitization for cultural heritage objects, a black background is used in order to obtain a large deviation to the color of the object and to prevent reflections. However, there are objects in cultural heritage that tend to be darker and can be difficult to mask with such a setup. In addition, some cultural heritage scans may require the use of mounts that often are of different colors. Binary masks are created segmenting the scanned object from the background using the captured image set as input to guarantee the digitization of the object only. This data is used to reconstruct the first intermediate 3D model. With this 3D model new camera positions are calculated to guarantee that meaningful images of the object are captured for the next iteration step. This will be repeated until a sufficient quality of the 3D model is achieved. The automated iterative scanning process only works effectively if the intermediate 3D models are complete and do not contain noise around the object.

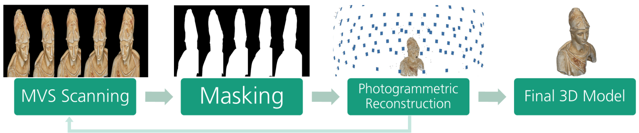


Figure 1: *Capturing an image set of the object is the first step in our automated iterative scanning workflow. The masked images are reconstructed resulting in a 3D model. Based on the 3D model new images are acquired, which are used again for a reconstruction until a certain quality is reached.*

Currently two different methods are used to calculate the masks: Threshold-based (TBM) and Color-based (CBM) image masking. The TBM compares each pixel value with a given threshold value. If the pixel value exceeds the threshold, the pixel will be removed, otherwise it will be kept. This only works well for objects with a high deviation in brightness and color to the background and the mount. This can be the case in a controlled environment, including TBM in this comparison as a baseline. Instead of defining just one threshold value, the CBM is using several Gaussian functions fitted by Gaussian Mixture Models defining the color of the object and the background, followed by an extraction using GrabCut [BTK*20]. Additional input in form of marked pixels is required for the object and the background, generated manually by a user. This must be done only on some images of the whole image data set to create a color segmentation model, that segments all images automatically. Problems can occur with this segmentation approach if the color of the background and the mount resembles the color of the object. The calculated masks have a significant impact on the scanning process shown in Figure 1. Masking out too much can lead to missing information in the images and thus to a false reconstruction of the intermediate 3D model. Additionally, noise in form of reconstructed background or mount is provoked if the masks contain additional information besides the object, e.g. an extension of the silhouette information by a bounding box. Both cases lead to an inefficient or abortive scanning workflow, generating too less or too many new camera positions, sometimes even positions that does not contain the object.

We therefore present a more generalized solution using several AI-based approaches to avoid parameter tuning of the TBM and CBM, and manual user input. Two models for single image segmentation, the U-Net and the Segment Anything Model (SAM), as well as two state-of-the-art models, segmenting the same object in multiple images (co-segmentation), the Unified Framework for Co-Object segmentation (UFO-Net) and the Group Collaborative learning Network (GCO-Net+), are tested.

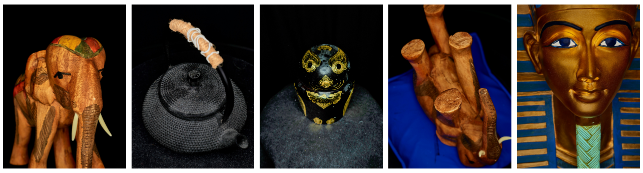


Figure 2: *Objects that are analyzed from left to right: An Elephant, a Teapot, an Owl, an Elephant stabilized on a blue mount and a Tutankhamun replica.*

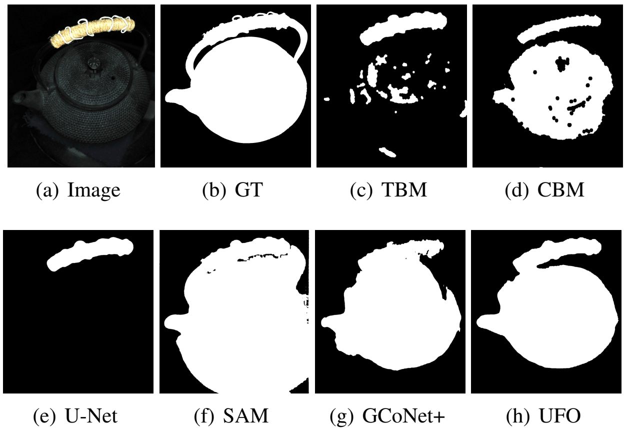
U-Net One of the first touch points of AI based image segmentation was the U-Net, which has been used within biomedicine to segment cells on light microscopy images [RPB15]. A latent vector is trained using a single image based fully connected convolutional network structure, whereupon the input images from the latent vector are estimated using deconvolution to generate the masks.

SAM SAM, as introduced by Kirillov et al. [KMR*23], has been trained on a dataset of 11 million images and 1.1 billion masks, and has strong zero-shot performance on a variety of segmentation tasks. The model consists of a heavyweight image encoder which outputs an image embedding that is then queried by a variety of input prompts to produce object masks at real-time speed.

UFO-Net Yukun et al. introduced the UFO-Net [SDS*22] for object co-segmentation and video salient object detection. It is structured in its base as a classical encoder decoder network using convolutional neural networks for feature extraction. In order to extract the dependencies between features of multiple images, it makes use of the recently introduced visual transformers and their self-attention mechanism.

GCoNet+ GCoNet+ [FFF*21] is another network for co-segmentation object detection. Most networks focus only on the similarities between objects of the same group in several images. This network also trains the difference to other image groups in order to better identify the actual object. As with the other networks, this intra and inter image group comparison happens on feature representations inside an encoder decoder structure.

All mentioned networks, except for SAM, whose pretrained version is tested, are trained end to end in a supervised manner. Thus, it is necessary to create ground truth masks for all training images, originating from real scans. This dataset consists of 16 different objects with ≈ 4000 images and hand-made masks. If available, we used the pretrained weights given by the authors as a starting point for training. All training was done on an Intel Xeon CPU E5-2687W @ 3.10 GHz, 256 GB RAM, GeForce RTX 2070 Super 8GB VRAM.

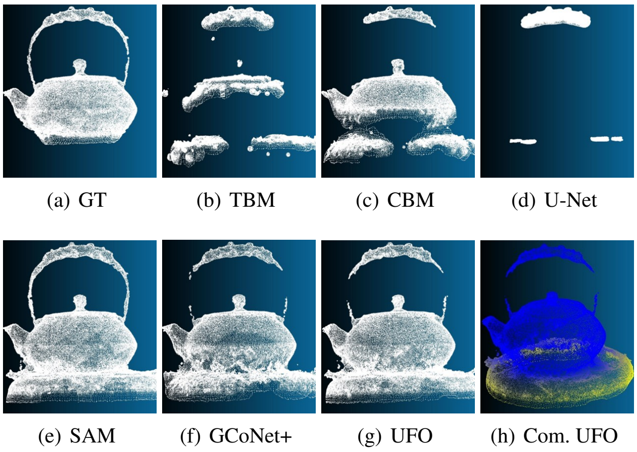

(a) Image (b) GT (c) TBM (d) CBM

(e) U-Net (f) SAM (g) GCoNet+ (h) UFO

Figure 3: *Exemplary masks of the Teapot show how the different methods cope with the problem of masking a dark object.*

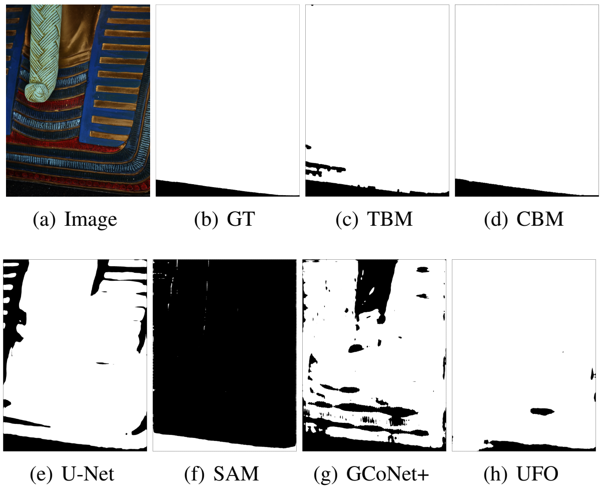
3. Results and Conclusion

For the comparison of the mentioned masking methods, scans with five different objects were performed. The scanning setup corresponds to a mass digitization with regard to the background and mount as described in Section 2. According to the cultural heritage domain the objects for scanning were chosen and are shown in figure 2. The Elephant figure has been successfully scanned before using the conventional methods (TBM and CBM) and serves as a reference. It was scanned a second time supported by a blue mount to recreate a common difficult scenario. The Teapot and the Owl represent the cases where it is difficult to distinguish between object and background. Finally, the Tutankhamun replica was deliberately chosen as a larger object to observe how the AI methods behave when some images do not contain any background. For the comparison between the conventional and AI-generated masks, ground truth (GT) masks were created for the mentioned objects using Adobe Photoshop. Figure 3 shows examples for the masks created with the different methods. To check the quality of the generated masks a 2D and 3D comparision is performed. In the 2D approach, all generated masks are compared with the GT masks in a pixelwise manner. The average accuracy (Acc.) and F_1 score for each method and all objects are shown in Table 1. The 3D comparison analyzes the effect of the masks within the scanning process. For this purpose, the scanned objects were reconstructed with the AI-generated, conventionally generated and GT masks. Again, the reconstructions with the different masks are compared by computing a cloud-to-cloud distance to the reconstruction that uses the GT masks [Clo]. From these distances, the mean 3D accuracy (3D Acc.) and completeness (Com.) are calculated and shown in Table 2. A small 3D accuracy value corresponds to a reconstruction of the object close to the ground truth and the completeness describes the number of correctly reconstructed points. Optimally, both metrics should be small, but for the automated iterative scanning workflow, a small 3D accuracy value is more important than a small completeness. If the 3D accuracy is small and the completeness high, too many points have been reconstructed, but no information has been lost. If the 3D accuracy is high and the completeness small, parts of the reconstruction are missing. All methods succeeded in creating good masks for the Elephant, the easy reference object. Regarding the scanning of dark objects GCoNet+ performs overall best. Of all AI-based methods it most successfully separates object and background, but often creates incoherent masks, especially if the object is as diverse in color as the replica of Tutankhamun, visible in Figure 5. The UFO shows the most consistent results with an overall accuracy of 94.2% and a F_1 score of


(a) GT (b) TBM (c) CBM (d) U-Net

(e) SAM (f) GCoNet+ (g) UFO (h) Com. UFO

Figure 4: *Pointclouds of the Teapot using different masks within the reconstruction. Image h) shows the completeness between the UFO and the GT model, with the correctly calculated points (blue) and points that are not included in the GT model (yellow).*


(a) Image (b) GT (c) TBM (d) CBM

(e) U-Net (f) SAM (g) GCoNet+ (h) UFO

Figure 5: *Exemplary masks of the Tutankhamun replica. It can be seen, that the SAM mask generated an inverse mask due to false object detection. In addition, the mask generated by the UFO network is more coherent than the mask from the GCoNet+.*

	Elephant		Teapot		Owl		Elephant + Mount		Tutankhamun		**All**	
Method	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1	Acc.	F_1
TBM	0.976	0.973	0.751	0.130	0.828	0.249	0.899	0.903	0.902	0.942	0.871	0.639
CBM	0.974	0.969	0.864	0.606	0.832	0.238	0.967*	0.965	**0.997**	**0.998**	0.927	0.755
U-Net	0.990	0.986	0.752	0.091	0.855	0.386	0.980*	0.978	0.738	0.834	0.863	0.655
SAM	0.925	0.918	0.783	0.697	**0.981**	**0.943**	0.896	0.899	0.163*	0.205	0.750	0.732
GCoNet+	0.962	0.954	**0.905**	**0.765**	0.945	0.816	0.922	0.878	0.747	0.823	0.896	0.847
UFO	**0.992**	**0.988**	0.854	0.679	0.936	0.782	**0.991***	**0.990**	0.936	0.960	**0.942***	**0.880***

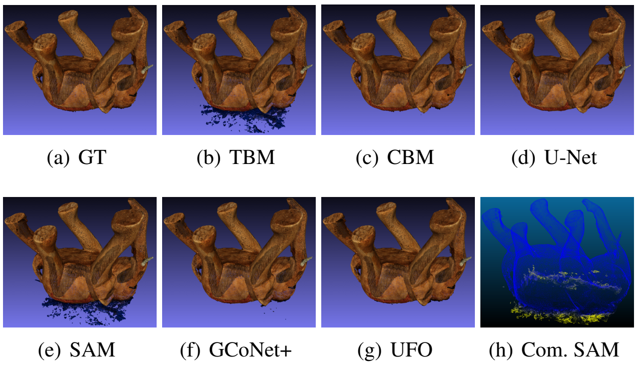
Table 1: *Average accuracy and F_1 score for each object and method.*

	Elephant		Teapot		Owl		Elephant + Mount		Tutankhamun		**All**	
Method	3D Acc.	Com.	3D Acc.	Com.	3D Acc.	Com.	3D Acc.	Com.	3D Acc.	Com.	3D Acc.	Com.
TBM	**0.224**	**0.210**	12.4	13.2	3.86	34.8	0.299	2.24	0.842	0.262	3.52	10.2
CBM	0.294	0.224	4.95	11.9	5.72	8.67	0.280	0.215	**0.171**	**0.170**	2.28	4.24
U-Net	0.259	0.271	44.0	**4.95**	3.96	1.71	0.266	**0.203**	1.11	0.265	9.92	**1.48**
SAM	0.243	0.277	**0.691**	12.3	**0.497***	**0.412**	0.258	2.21	29.5	2.63	6.23	3.55
GCoNet+	0.530*	0.289	1.32	12.9*	1.37	1.99	0.288	0.261	4.03	0.310	1.51	3.15
UFO	0.249	0.274	0.875	12.9*	1.61	0.656	**0.192***	0.211	1.15	0.276	**0.815**	2.86

Table 2: *Average 3D accuracy and completeness of the pointclouds reconstructed using the respective masks in millimeter [mm].*

88.0%. It generates overall decent results for all objects and correctly differs between object and background or mount. Only the masks of the dark objects, which tend to be too small, leave room for improvement. The blue mount, which is used to stabilize the Elephant could also be segmented successfully using the UFO network achieving an accuracy of 99.1%, followed by the U-Net with 98.0%. Looking at the conventional methods, the CBM reaches an accuracy of 96.7%, which makes it comparable to the UFO and U-Net. The results of the SAM are very mixed. If the network recognizes the object and can correctly distinguish it from the background, it produces the best masks of all methods. However, it often has difficulties capturing the entire object, especially if it takes up a large portion or the entire image, resulting in only 16.3% correctly classified pixels on average for the Tutankhamun replica. Similar to the 2D comparison the reconstruction of the Elephant is nearly the same independently of the used mask. Only the GCoNet+ masking achieves a higher mean 3D accuracy with 0.530mm than any other method. In general, the results of the 2D comparison are consistent with the 3D comparison. The SAM can segment the objects pretty good, if it can identify the object within the image. Looking at the mean 3D accuracy of the Owl and the Tutankhamun reconstructions in Table 2 this was not always the case. The UFO-based masking produces stable results independent of the masked object, followed by the GCoNet+. Incorrectly, the mount of the Teapot is also reconstructed leading to high mean completeness value of 12.9mm, as seen in Figure 4. Even though the completeness for the TBM, CBM and U-Net masking is lower, the reconstruction contains too less information for a successful automated iterative scanning workflow. Equivalent results are achieved with the Owl, where SAM has the lowest mean 3D accuracy of 0.497mm. Looking at the Elephant stabilized with the blue mount, all methods except for SAM and TBM segmented the mount successfully, as can be seen in Figure 6. The UFO network reaches the lowest mean 3D accuracy of 0.192mm. Lastly, the Tutankhamun replica could be reconstructed well with all masking methods, except for the SAM-based masks, concluding that UFO-, GCoNet+ and the U-Net-based masks do not have any problems with images without any visible background.

In summary, we can conclude that the AI masks benefit the reconstructions and thus the automated iterative scanning workflow, especially for objects with a similar color as the background. For simple objects, AI masks can be used without hesitation, however, the conventional methods have achieved good results for those objects as well. In conclusion, the UFO network provides the best results and the highest stability with the overall lowest mean 3D accuracy, followed by GCoNet+. U-Net is obsolete and not recommended since U-Net resembles conventional methods. SAM has a great potential for further improvements, if it is able to detect the object correctly.


(a) GT (b) TBM (c) CBM (d) U-Net

(e) SAM (f) GCoNet+ (g) UFO (h) Com. SAM

Figure 6: *Pointclouds of the Elephant placed on a blue mount using different masks within the reconstruction. Image h) shows the completeness between the SAM and the GT model, with the correctly calculated points (blue) and points that are not included in the GT model (yellow).*

References

[BTK*20] BUELOW M., TAUSCH R., KNAUTHE V., WIRTH T., GUTHE S., SANTOS P., FELLNER D. W.: Segmentation-Based Near-Lossless Compression of Multi-View Cultural Heritage Image Data. Eurographics Workshop on Graphics and Cultural Heritage: 978-3-03868-110-6 (ISBN), 2312-6124 (ISSN), Eurographics Association. 2

[Clo] CLOUDCOMPARE: (Version 2.12.04) [GPL Software] (2023). URL: http://www.cloudcompare.org/. 3

[FFF*21] FAN Q., FAN D.-P., FU H., TANG C.-K., SHAO L., TAI Y.-W.: Group collaborative learning for co-salient object detection. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 12283–12293. 2

[KMR*23] KIRILLOV A., MINTUN E., RAVI N., MAO H., ROLLAND C., GUSTAFSON L., XIAO T., WHITEHEAD S., BERG A. C., LO W.-Y., DOLLÁR P., GIRSHICK R.: Segment anything, 2023. arXiv:2304.02643. 2

[MGC13] MANFREDI M., GRANA C., CUCCHIARA R.: Automatic single-image people segmentation and removal for cultural heritage imaging. In *New Trends in Image Analysis and Processing–ICIAP 2013* (2013), Springer, pp. 188–197. 1

[RPB15] RONNEBERGER O., P.FISCHER, BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015), vol. 9351 of *LNCS*, Springer, pp. 234–241. 2

[SDS*22] SU Y., DENG J., SUN R., LIN G., WU Q.: A unified transformer framework for group-based segmentation: Co-segmentation, co-saliency detection and video salient object detection, 2022. arXiv:2203.04708. 2

[SRT*14] SANTOS P., RITZ M., TAUSCH R., SCHMEDT H., MONROY R., STEFANO A. D., POSNIAK O., FUHRMANN C., FELLNER D.: Cultlab3d - On the verge of 3D mass digitization, 2014. 1

[STD*20] SANTOS P., TAUSCH R., DOMAJNKO M., RITZ M., KNUTH M., FELLNER D. W.: Automated 3D Mass Digitization for the GLAM Sector. Archiving 2020 online: IS&T. - 2161-8798 (ISSN) 2168-3204 (E-ISSN). - (2020) (Archiving Conference), IS&T. 1

[ZZX*20] ZHANG J., ZHOU Y., XIA K., JIANG Y., LIU Y.: A novel automatic image segmentation method for chinese literati paintings using multi-view fuzzy clustering technology. *Multimedia Systems 26* (2020), 37–51. 1

Proceedings published by Eurographics - The European Association for Computer Graphics.