International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. 15, No.1 (Mar-24)

http://dx.doi.org/10.12785/ijcds/150194

Efficient 3D Instance Segmentation for Archaeological Sites Using 2D Object Detection and Tracking

Maad kamal Al-anni[1,2] **and Pierre DRAP**[2]

[1]*Computer Engineering Department, College of Engineering, Al-Iraqia University, 7366 Baghdad, Iraq.*
[2]*Aix Marseille University, CNRS, ENSAM, Universit´e De Toulon, LIS UMR 7020, 13397 Marseille, France.*

Received 26 Jan. 2024, Revised 3 Mar. 2024, Accepted 5 Mar. 2024, Published 10 Mar. 2024

Abstract: This paper introduces an efficient method for 3D instance segmentation based on 2D object detection, applied to the photogrammetric survey images of archaeological sites. The method capitalizes on the relationship between the 3D model and the set of 2D images utilized to compute it. 2D detections on the images are projected and transformed into a 3D instance segmentation, thus identifying unique objects within the scene. The primary contribution of this work is the development of a semi-automatic image annotation method, augmented by an object tracking technique that leverages the temporal continuity of image sequences. Additionally, a novel ad-hoc evaluation process has been integrated into the conventional annotation-training-testing cycle to determine the necessity of additional annotations. This process tests the consistency of the 3D objects yielded by the 2D detection. The efficacy of the proposed method has been validated on the underwater site of Xlendi in Malta, resulting in complete and accurate 3D instance segmentation. Compared to traditional methods, the object tracking approach adopted has facilitated a 90% reduction in the need for manual annotations, The approach streamlines precise 3D detection, establishing a robust foundation for comprehensive 3D instance segmentation. This enhancement enriches the 3D survey, providing profound insights and facilitating seamless exploration of the Xlendi site from an archaeological perspective.

Keywords: Underwater archaeology, AI, Convolutional Neural Network (CNN), 3D Instance Segmentation, and Underwater photogammetry.

1. Introduction

The approach developed in this article has been experimented with during the deep-sea archaeological excavation of Xlendi in Malta, led by Prof. Timmy Gambin. The initial photogrammetric surveys date back to 2009, and a collaboration between two laboratories from Aix-Marseille University—the LIS (Laboratory of Computer Science and Systems) and the CCJ (Centre Camille Julian)—along with the University of Malta, has facilitated annual excavation and photogrammetric survey campaigns over a three-year period [1]. The excavations and 3D surveys documenting the evolution of the excavation site continued until 2022, under the direction of the University of Malta. A view of the surface layer, composed of amphorae and a grinding stone obtained through photogrammetry in 2014, is illustrated in Figure 1.

Figure 1. Orthomosaic image, generated by photogrammetry on the Xlendi wreck, showing amphorae and grinding stone laying on the sediment.

In this paper, we utilize a 2D-to-3D link to transfer 2D object information — specifically amphorae in our case, as shown in Figure 2 — from the given dataset. Furthermore, given that we have a complete 3D model of the site obtained through photogrammetry, we establish a link between the detected objects in the images and their instance segmentation in the 3D model. This link is a 2D-to-3D correspondence that allows for the integration of information across these dimensions.

A traditional object detection approach, based purely on

E-mail address: muad.rashed@aliraqia.edu.iq, pierre.drap@lis-lab.fr

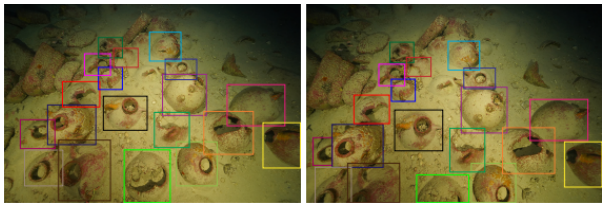

Figure 2. Example images for the underwater site showing artifacts (amphorae) in 2D, that need to be detected automatically.

color or spatial information, may fail, as stated by [1]. Recent advances in machine learning, and more specifically in deep learning, have led to robust end-to-end object detection methods. By leveraging a bidirectional 2D-to-3D link and employing multitask learning, 3D instance segmentation can be accomplished using a 2D object detection approach; or at least, a close part of the 3D model can be identified. This is achieved by incorporating instance segmentation into 2D object detection.

To build a robust model, it is necessary to label the entire dataset and prepare it for the training phase. Given the size of the dataset, which comprises over 30,000 images from surveys conducted over the past 10 years, manually labeling the entire dataset poses a significant challenge. To address this, we adopt a new approach that mitigates this issue by using an ad-hoc evaluation process, which can be implemented if further annotations are needed. Ultimately, the need for manual labeling of the entire dataset is reduced by 90% compared with traditional techniques, thanks to object tracking.

The Metashape software has been utilized for photogrammetric calculations since 2009. Specific scripts are developed in Python to manage the projections of 2D labeled instances to obtain a probability area of amphorae presence in 3D space. This allows for the precise identification of labels found on photos with the 3D instances of amphorae. The result is a 3D model that focuses on the visible part of an amphora or a fragment thereof, including a small part of the sediment and a visible portion of the amphora. Future work will lead us to develop proposals regarding the determination of the typology of isolated amphorae.

In simple terms, this research leverages the 2D object detection method (YOLOV4) in two key stages to achieve Efficient 3D instance segmentation improves archaeological site exploration comprehensively. First, we employ an initial training phase to minimize human intervention in labeling the entire dataset. This is achieved through the use of a semi-automatic image annotation method combined with a novel ad-hoc evaluation process. Subsequently, once the dataset is labeled, we proceed with the final training phase to obtain a robust model, which serves as the input to the second stage. In the second stage, we address this challenge by adapting YOLOV4 to the 2D/3D linkage facilitated by photogrammetry. For both implementations, we rely on open-source software, one for labeling purposes and another for 3D instance segmentation using YOLOV4.

The structure of this paper is as follows: Firstly, we introduce the problem in the context of archaeological sites and highlight our main contributions. Next, we review related work and explain how our approach differs from others. The subsequent section outlines our methodology, which involves a semi-automatic image annotation method combined with object tracking, utilizing the YOLOV4 model as our reference. Following this, we describe the process of 3D reconstruction from 2D images, bridging the gap between 2D object detection and 3D projection using photogrammetry. Finally, we engage in a general discussion about the final implementation, its benefits, and potential avenues for future work.

2. Related Works

Although 3D instance segmentation methods have seen notable advancements, as shown by previous research [2][3][4], these approaches have been limited to a few models. In other words, the detection and classification of 3D objects is only relevant, reliable, and usable in a truly experimental context when the training of object detection methods is done in 2D. This is similar to the problem at the heart of this work where, fortunately, 2D and 3D data are intimately linked and the 2D data are very abundant.

Takmaz et al.[5] introduced OpenMask3D, a zero-shot approach for open-vocabulary 3D instance segmentation. Their approach addresses the limitations of existing methods and enables segmentation of object instances based on free-form queries. The experimental results demonstrated its superior performance compared to other approaches.

Rozenberszki et al.[6] proposed UnScene3D, an unsupervised approach for class-agnostic 3D instance segmentation. The method generates pseudo masks using self-supervised features and refines them through self-training. The method outperforms existing approaches by over 300% in terms of average precision, even in challenging scenes.

Kontogianni et al.[7] proposed an interactive approach for 3D instance segmentation, allowing users to collaboratively segment objects in 3D point clouds. Unlike fully supervised methods, this approach does not require costly training labels and adapts to new environments. Users can directly interact with 3D point clouds, clicking on objects of interest to achieve accurate segmentation with minimal effort. This method opens up possibilities for applications in Augmented Reality (AR)/Virtual Reality (VR) and human robot interaction, facilitating efficient labeling of diverse 3D datasets.

Chibane et al.[8] presented a weakly supervised approach for 3D semantic instance segmentation using bounding box labels. The method, called Box2Mask, incorporates a deep model based on Hough voting and a specialized clustering method. It achieves competitive performance on the ScanNet test dataset and demonstrates successful instance segmentation on the ARKitScenes dataset using only bounding box annotations.

A fully-convolutional 3D point cloud instance segmentation method avoids clustering and its associated challenges [9]. It utilizes per-point prediction and optimal transport for target assignment, achieving promising results on the ScanNet and

S3DIS benchmarks. The method removes inter-task dependencies and provides a simple and accurate 3D instance segmentation framework.

Zhong et al.[10] addressed the challenging task of 3D instance segmentation by proposing a novel framework. The approach involves learning offset vectors for points and grouping them using a hierarchical point grouping algorithm. Multiscale groups were used for instance prediction, and MaskScoreNet was employed to refine the segmentation results. Experimental results on the ScanNetV2 and S3DIS benchmarks demonstrated the model's performance, achieving a 66.4% mAP with a 0.5 IoU threshold on ScanNetV2, outperforming the state-of-the-art method by 1.9%.

Shen and Stamos[11] proposed a novel object segmentation and detection system that utilizes 2D detection to generate frustums, followed by a 3D convolutional-based method, Frustum VoxNet, for 3D instance segmentation and object detection. The system achieves fast 3D inference with RGB-D images and comparable accuracy with depth-only images, making it suitable for low-light conditions or RGB-absent sensors. The inclusion of segmentation in the pipeline improves detection accuracy while providing 3D instance segmentation.

Shao et al.[12] presented an RGB-D based instance-level segmentation method that provides detailed information about object location, geometry, and quantity, which are crucial for safe decision-making in real-world environments. The model uses object occupancy moments to represent instances and an hourglass DNN for 3D position, size, and pose voting. Clustering and object-centric training achieve superior performance compared to the state-of-the-art on both synthetic and real-world datasets.

All approaches face challenges with large-scale 3D datasets, such as occlusions, noise, and varied object shapes and sizes, making annotating 3D data for training laborious. Additionally, 3D instance segmentation techniques may struggle with objects having complex boundaries, leading to less accurate results. Our method tackles these issues with semi-automatic image annotation and a unique evaluation process. Leveraging the bidirectional link between 2D and 3D via photogrammetry, we isolate point clouds of interest like amphorae from a 2D perspective. This innovative approach streamlines 3D instance segmentation while reducing labeling effort by 90%.

3. Methodology

A. *Object Detection Approaches*

Object detection has emerged as one of the most complex fields in computer vision, undergoing significant advancements over the past decade. In essence, this technique initially aims to: 1) identify the spatial location of objects within an image, known as object localization, and 2) assign them to specific categories, known as object classification. Algorithms vary in efficiency and scalability. Examples include Region-Based Convolutional Neural Networks (R-CNN)[13], Fast R-CNN[14], Faster R-CNN[15], Histograms of Oriented Gradients (HOG)[16], the Region-based Convolutional Network method (R-CNN)[13] again listed here, Region-based Fully Convolutional Networks (R-FCN)[17], Single Shot Detector (SSD)[18], Spatial Pyramid Pooling (SPP-net)[19], and You Only Look Once (YOLO)[20].

In the context of 3D surveys made from 2D images, the YOLO approach (fully 2D) is considered advantageous due to its real-time processing speed. This makes it suitable for end-to-end applications that require rapid object detection. YOLOV4 divides the output space of bounding boxes into a grid of default boxes with different aspect ratios and scales per feature map location. When making predictions, the network assigns scores to each object category within each default box and refines the box to better match the object's shape. Furthermore, the network combines predictions from multiple feature maps at different resolutions to effectively handle objects of various sizes. The YOLOV4 architecture is illustrated in Figure 3.

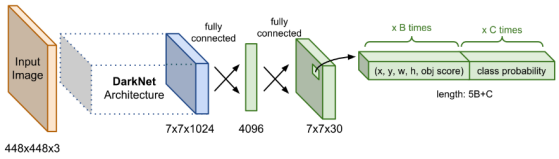

Figure 3. DarkNet Network of YOLOV4 Architecture.

Given that the complete detection process is incorporate within one network, it is capable of undergoing comprehensive optimization based on detection performance, YOLOV4 detection network has 24 convolutional layers, followed by 2 fully connected layers. Alternating convolutional layers reduce the feature space from preceding layers. This approach has undergone several improvements since its inception. For instance, recent versions have become considerably more accurate and many times quicker than the original.

B. *Semi-Automatic Image Annotation Coupled with Object Tracking Method*

The main challenge we encounter is managing a vast dataset consisting of more than 30,000 images covering surveys conducted over an extensive period, figure 4 illustrates the heuristic approach we employed to manage human intervention efforts for annotating this extensive dataset. To address this, we utilise a semi-automatic image annotation approach that leverages object tracking. Although a significant number of images still require manual labelling, the sequential nature of the application reduces the overall manual labelling workload significantly. Our approach performs similarity measure between consecutive images and selects one image per group of similar images. Only selected images are required to be labelled by extending an open source software. Then, an initial training phase is performed using this subset of images, menu buttons for detection and tracking as shown

in Figure 5.

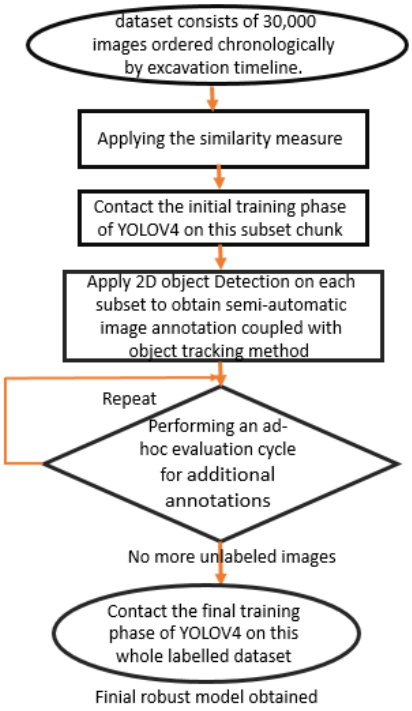

Figure 4. demonstrate the reduction of human intervention efforts by employing 2D object detection (YOLOV4) to handle the entire dataset of 30k images. This is achieved through a semi-automatic image annotation method coupled with a novel ad-hoc evaluation process, resulting in the acquisition of a robust final model.

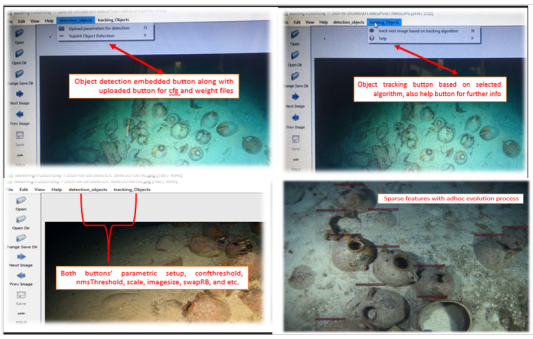

Figure 5. semi-automatic image annotation coupled with object tracking method.

The obtained model is used to automatically detect amphorae in the remaining unlabelled images, the detection process is refined using the results of a sparse feature point matching approach applied to an ad hoc evaluation process to labelled and unlabelled images within each labelled and unlabelled group if any. A second training phase is performed using all images to obtain the final robust model as shown in Figure 6.

The model utilizes LabelImg[1], an open-source graphical image annotation tool. This tool is indispensable for drawing bounding boxes around amphorae in 2D images and adding corresponding labels. LabelImg, compatible with both PyQt4 and PyQt5, facilitates a semi-automatic image annotation approach. Additionally, its adaptability extends to various formats. Notably, DarkNet[2] is aligned with the YOLOV4 format, optimizing it for datasets in our proposed model and related object detection tasks. Given its convenience and user-friendly interface, LabelImg is highly regarded in the deep learning community.

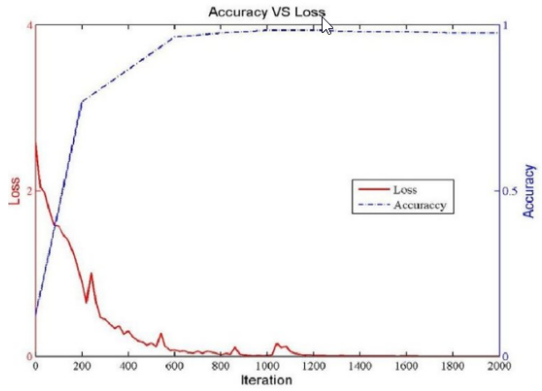

Figure 6. The accuracy and loss function graphs depict the performance of the finalized robust model, achieved by setting the maximum batches parameter to 2000 multiplied by the number of classes and IOU threshold = 0.5.

C. Link with Photogrammetry

Photogrammetry is widely recognized as the simplest and most effective method for conducting a 3D survey in an underwater environment. Requiring only a brief period for fieldwork, which is particularly advantageous in underwater contexts, this technique proves to be extremely beneficial, especially under the challenging conditions found underwater. Indeed, visibility underwater is severely limited, necessitating that photographs be taken in close proximity to the subject to minimize the water column between the object and the camera lens. This requirement leads to a proliferation of photographs and their enlargement. On the one hand, these factors complicate the photogrammetry process by increasing the number of images; on the other hand, they amplify the utility of the 2D documents, which we are already adept at exploiting. Moreover, we can exploit the sequential shooting mode, capturing images along a trajectory at regular intervals, a condition that is conducive to employing tracking techniques.

Consequently, we are able to develop a 2D-focused approach for the thousands of photographs and integrate these methods with the 3D model generated through

[1]https://github.com/HumanSignal/labelImg
[2]https://github.com/AlexeyAB

photogrammetry.
Photogrammetry computations are performed using the Metashape software[3], which provides the position of each photograph as well as a sparse 3D point cloud that encompasses the scene. These points, essential for calculating the pose of each photograph, are not intended to finely detail the scene. Nevertheless, each point is associated with the various photographs in which it has been detected.
As is now evident, the detection of artifacts in images by automated approaches, as proposed by well-known software suites like YOLOV4, can be projected into the 3D space delineated by photogrammetry. While a robust and effective method like YOLOV4 may not operate directly in 3D, we can still implement this type of approach by utilizing the 2D/3D linkage provided by photogrammetry.

D. Toward 3D Instance Segmentation Using YOLOV4

We use the bidirectional relationship between 3D and 2D reconstruction, which is at the core of the photogrammetry process, to link the 3D model with the 2D recognition approach For further illustration, please refer to the Figure 7.

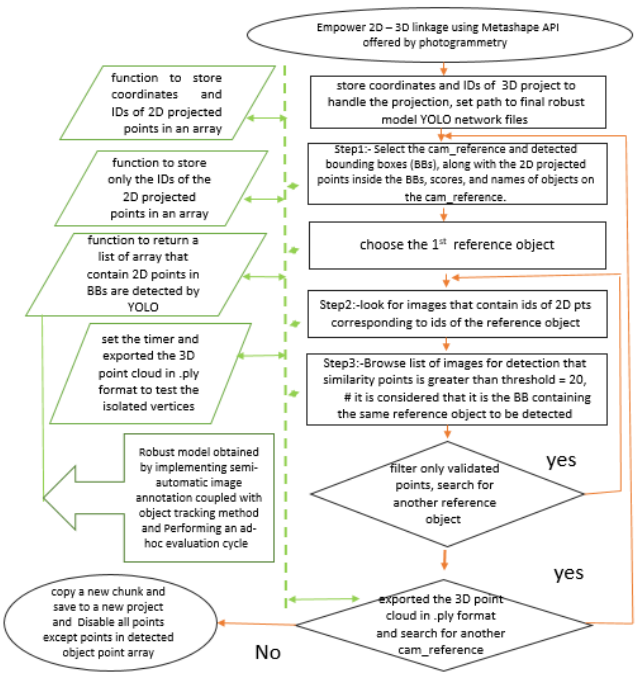

Figure 7. illustrate the application of a 2D-3D bidirectional link to operate a 2D object Detection Model, aiming to achieve 3D instance Segmentation using YOLOV4. This is done through projection with an extension of photogrammetry.

Contrary to 2D instance segmentation, 3D instance segmentation provides a more intuitive and informationally rich perspective. In a 2D image, the projections of 3D objects can overlap, complicating the differentiation of objects. In contrast, 3D space naturally separates different objects, offering clearer demarcation, yet object detection in 2D has proven to be robust and effective. On this basis, we introduce a specialized 2D-to-3D instance segmentation technique that leverages 2D object detection, specifically designed for the unique context of 3D archaeological scenes. Our approach enables precise 3D detection and establishes a solid foundation for 3D instance segmentation, which has enriched the 3D survey by providing comprehensive insights and a seamless exploration of the Xlendi site from an archaeological perspective.
The Metashape software serves to define spatial properties and utilizes a sparse 3D point cloud for the comprehensive representation and manipulation of objects within the scene. It is instrumental in determining the 3D scene's position and orientation based on a foundational 3D point cloud. These points, crucial for calculating photo orientations, may lack intricate scene details but are connected to the specific 2D photographs where the measurements were taken.
It is noteworthy that robust 2D object detection approaches like 2D object detection(YOLOV4) are not inherently designed for 3D applications. However, in our methodology, we bridge this gap by adapting YOLOV4 to the 2D/3D linkage provided by photogrammetry. In the context of the archaeological site Xlendi, our model identifies the intersections of the 3D tie-point cloud projections using this linkage, assigning them a valid status. This methodology diverges from conventional 3D instance segmentation methods that rely on voting and grouping of 3D orthocoordinates based on a clustering paradigm, which proves to be less effective for large and irregularly shaped objects.[21] [22] [23].
By contrast, our approach does not depend on any hand-tuned, distance-based clustering. Instead, all instances in the scene are simultaneously represented as Bounded Boxes (amphorae) with lower-left and upper-right 2D coordinates,
Point clipping in 2D involves determining whether a point lies within a specified BB or not, as shown in Figure 8 and Figure 9.
Let's consider a simple BB as an example. The rectangular region of BB is defined by its minimum and maximum coordinates (xmin, ymin) and (xmax, ymax). A 2D point (x, y) is cropped if it satisfies the following conditions:
$xmin \leq x \leq xmax$
$ymin \leq y \leq ymax$
If these conditions are met, the 2D point lies within the BB and is visible; otherwise, it is outside BB and needs to be invisible.
The cropped boxes mentioned are generated from a 2D trained model, previously identified as a robust model. Subsequently, our approach selects only the 3D maps of identified id_objects to identify intersections among 3D tracking correspondences. This is achieved by

[3]https://www.agisoft.com/

utilizing the 2D-3D bidirectional link, which is further extended by photogrammetry, as illustrated in Algorithm 1.

Algorithm 1: 3D Point visibility Algorithm

Data: List of all Point cloud $P(P_x, P_y, P_z, P_{track\_id})$, containing the points of the reference object $reduce\_id\_obj\_3D$

Result: Disable all points except points in detected object point array

for *each point pts in P* **do**

  if *not* pts_{track_id} *in* $reduce\_id\_obj\_3D$ **then**

    do $pts.valid = False$

  Data: copy a new chunk and save to a new project

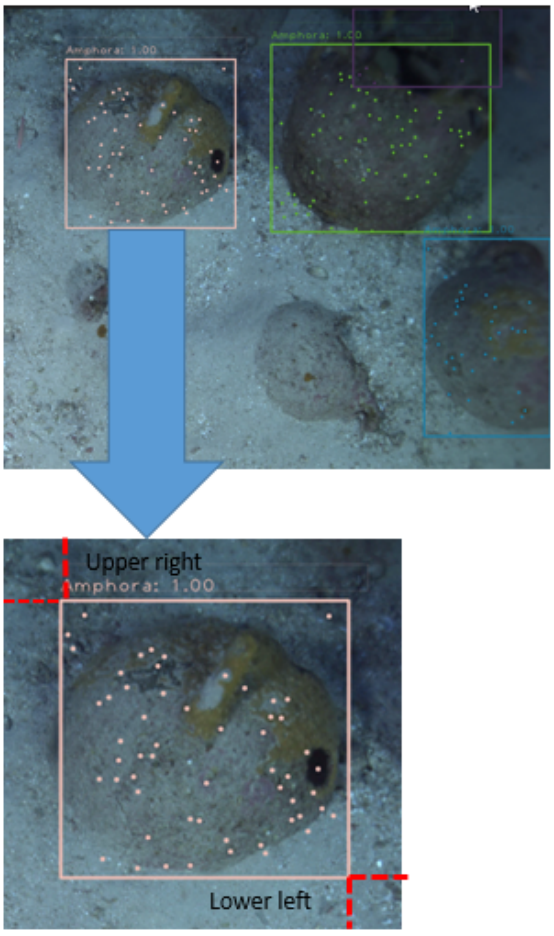

Figure 8. bounded box (amphora) with lower-left and upper-right 2D coordinates

The high confidence levels across the entire extent of instances lead to correct predictions. Different from modern approaches that predominantly rely on geometric clustering techniques [24] [25], our proposed approach does not require 3D-specific components, such as centre voting or manually tuned distance-based clustering.

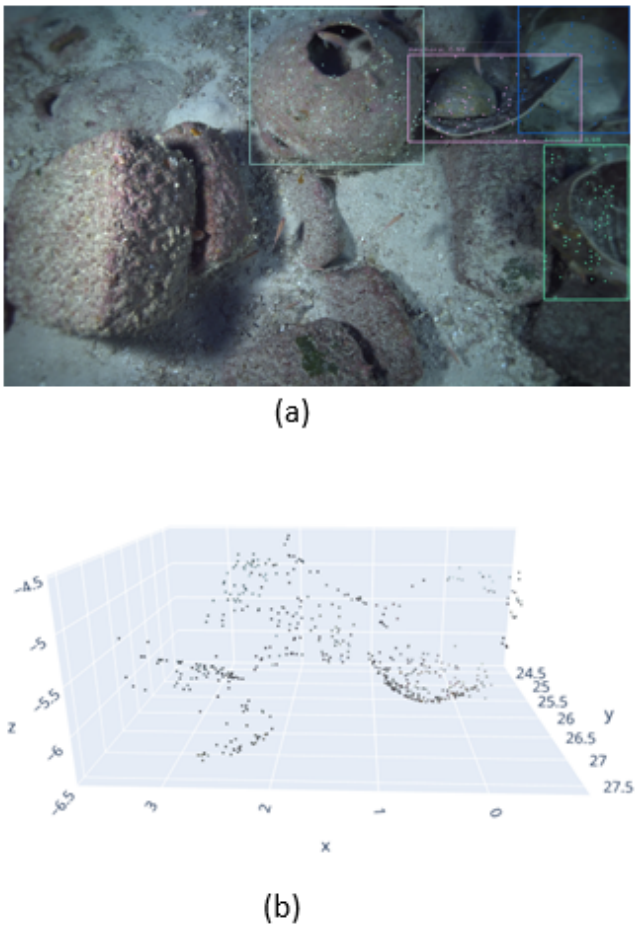

Figure 9. Result for camera reference = 68 (a) shows the bounded boxes for 4 amphorae in 2D with different coloured demarcations, (b) 3D tie-point plotting for four boxes.

on the basis of the success of recent intersection bounded-box detection and 3D id-objects accumulation through bounded-box voting and grouping with the utilisation of 2D-to-3D bidirectional links, our overall 3D instance segmentation using 2D object detection yields promising results across a various of challenging archaeological sites and heritage documentation tasks, exploring an ancient shipwreck off Xlendi Bay, Gozo, is now publicly available on the Google Play Store app.

For optimal processing of large sparse point clouds in 3D instance segmentation and their bidirectional links, we utilised the Metashape Python API. Our hardware setup includes a CPU with SSE 4.2 support, a high-end NVIDIA GPU, a recommended 16 GB RAM, and an

SSD for enhanced processing speed. On the software side, we ensure compatibility with Windows 10 and use Python version 3.8. Meeting these specifications guarantees efficient script performance and smooth API integration.
For instance, Camera Reference 66 which contains four amphorae with an id_objects' map of 590, requires roughly 47 s to processes. Figure 10 illustrates the relationship between the size of the id_objects' map and the associated processing time for each camera reference. Significantly, processing duration is largely dependent on the specific processors and graphics card in use.

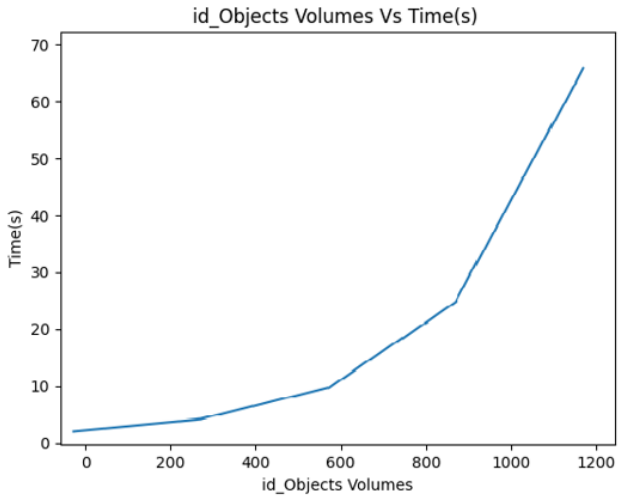

Figure 10. Relationship between the volume of id_objects' map and time(s) for given camera references.

A pivotal step is the amphorae segmentation within the observed scene, especially when preserving the id_objects' map for every camera reference in the scanned 3D project via the software. Subsequently, certain points are toggled to "false" validation, whereas others within the id_objects' map maintain their original validation status. This sets the stage for the project's concluding phase.
After adjusting the tie-point cloud, the subsequent phase entails analyzing overlapping images, pulling millions of points to craft a comprehensive 3D depiction of the observed scene. This cloud, capturing intricate geometries, lays the groundwork for mesh formation. The meshing stage transitions these points into a seamless 3D amphora model, encompassing validated vertices, edges, and faces, delineating segmentation structures. after mesh formation, texturing commences[26], primary images overlay this mesh, mapping intricate colors and patterns from the photos onto the 3D amphora facade. The outcome is a lifelike 3D model, reflecting precise geometry and visual veracity.
Upon refining the tie-point cloud, the process progresses to meshing and texturing by evaluating overlapping imagery. This final stage harnesses vast data points, generating a comprehensive 3D representation of the surveyed scene. The detailed point cloud lays the groundwork for the creation of the mesh. Meshing then transforms these aggregated points into a seamless 3D amphora structure, characterized by vertices, edges, and faces, providing a sophisticated segmentation blueprint. Following the creation of the mesh, attention shifts to texturing. In this phase, original images are superimposed onto the mesh, transferring intricate colors and patterns onto the 3D amphora surface. The end result is a photorealistic 3D model, notable for its geometric precision and visual fidelity [27][28].

4. Upcoming Investigations

The progress in 3D instance segmentation using photogrammetry significantly enhances the surveying and digitization of the Xlendi site, offering valuable insights for archaeologists. Despite advancements in automation, data acquisition, and processing speeds, deep-sea archaeological sites pose challenges like extreme pressure, limited visibility, and artifact preservation. Specialized equipment and methods are crucial for exploration, making research and recovery complex. Ongoing research focuses on semantic segmentation in underwater environments and temporal monitoring. Despite these challenges, acquiring digital models for dissemination has become more accessible, especially for those proficient in photography and basic surveying principles.
The significance of Virtual Reality (VR) in archaeology and cultural heritage spans over three decades, with VR techniques evolving to offer immersive experiences within archaeological sites. These advancements not only reproduce the present state of heritage but also enable the simulation of the past, a concept termed cyberarchaeology. Various projects exploring the application of Virtual and Augmented Reality in underwater locations aim to facilitate virtual exploration for non-divers, raise awareness, and advocate for Underwater Cultural Heritage (UCH) through educational games. Additionally, these technologies support in-depth examination, analysis of complex excavations, and monitoring their development over time[29] [30] [31] [32] [33] [34].
The proposed 3D model, relying on 3D instance segmentation, is designed to consist of two distinct backend components and a unified frontend. Device sensors facilitate localization and mapping, enabling seamless integration. The first backend integrates ontology and the Metashape API for the Xlendi archaeological study, utilizing photos collected since 2004. The second backend visualizes site geometry by segmenting amphorae in VR/AR.
In addition to offering a comprehensive view of the 3D model, it provides a non-invasive technique, allowing us to visualize various stratigraphic structures and the relationships between different stratigraphic units (US). It localizes the object of interest based on the user's movements and adjusts the point of view of the device within the 3D model. The pilot sample demonstrates its

effectiveness and efficiency, particularly when intuitively examining the 3D tie-point cloud. Ongoing research aims to validate its performance for the 3D dense cloud, including mesh and texture, Future work will lead us to develop proposals regarding the determination of the typology of isolated amphorae, and Future improvements could potentially enhance run-time efficiency in 3D scene, especially when employing recent version of YOLO on 2D images, with or without an enhanced labelling process.

5. Conclusion

In this paper, we introduce a novel, precise, and efficient algorithm for optimally processing large sparse point clouds in 3D instance segmentation, including their bidirectional links. We develope a 2D-based 3D detection system using a 2D object detection approach. Although YOLO V4, as implemented by DarkNet, solely performs detection, our implementation observations reveal the recent advent of YOLO V8 panoptic, which performs both segmentation and detection on 2D images, offering enhanced accuracy. Through the integration of 2D object detection using YOLO V4 for 3D instance segmentation, we gain a comprehensive understanding of 3D scenes, encompassing sparse and dense points.

Our detection success is showcased by achieving 85-87% performance, based on ad hoc labelling technique and 2D object detection by YOLO V4, decreasing labelling effort by up to 90%. Our model validates its hypothesis of leveraging a bidirectional 2D-to-3D link and employing 2D-to-3D segmentation based on 2D object detection approach by incorporating an instance segmentation and an object detection with reasonable accuracy, it provides comprehensive insights and a smooth exploration of the Xlendi site from the perspective of archaeologists.

Future improvements could potentially enhance run-time efficiency in 3D scene, especially when employing recent version of YOLO on 2D images, with or without an enhanced labelling process. The use of labelled datasets in supervised neural networks significantly affects the overall performance. Furthermore, our capability to enhance the accuracy rate directly results in a reduction of false positives and false negatives, thereby improving the feature extraction of target objects.

Our flexibly designed labelling process can operate in data-centric AI with the image labeling tool in High Resolution (HR) and Low-Level Radiometry (LLR) for smart cropping techniques. It is also expand the volume of a trained dataset, the use of a semi-automatic image annotation coupled with object tracking method and an ad hoc evaluation process are applicable. We foresee the application of our techniques in real-time robotics applications. One potential direction for future research is the incorporation of our system into a robotic platform.

Acknowledgments

I acknowledge the support received from the Ministry of Higher Education and Scientific Research (MoHESR), Republic of Iraq, Al-iraqia University, College of Engineering. I also thanks the Embassy of France in Iraq (Campus France), which granted a Scholarship, furthermore, I appreciate all the participants of this work and the Axi-Marseille University, LIS laboratory members.

Furthermore, this project is based on a long-standing cooperation, since 2009, between the University of Malta and Aix-Marseille University. Then, the University of Malta, under the direction of Prof. T. Gambin, conducted the excavations of the Xlendi wreck alone at a depth of over 100 meters.

References

[1] P. Drap, D. Merad, B. Hijazi, L. Gaoua, M. Nawaf, M. Saccone, B. Chemisky, J. Seinturier, J.-C. Sourisseau, T. Gambin, and F. Castro, "Underwater photogrammetry and object modeling: A case study of xlendi wreck in malta," *Sensors*, vol. 15, pp. 30 351–30 384, 12 2015.

[2] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," 2019.

[3] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3d: Mask transformer for 3d semantic instance segmentation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8216–8223.

[4] H. Thomas, C. R. Qi, J. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," *CoRR*, vol. abs/1904.08889, 2019. [Online]. Available: http://arxiv.org/abs/1904.08889

[5] A. Takmaz, E. Fedele, R. W. Sumner, M. Pollefeys, F. Tombari, and F. Engelmann, "Openmask3d: Open-vocabulary 3d instance segmentation," *CoRR*, vol. abs/2306.13631, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.13631

[6] D. Rozenberszki, O. Litany, and A. Dai, "Unscene3d: Unsupervised 3d instance segmentation for indoor scenes," in *ArXiv Preprint*, 2023.

[7] T. Kontogianni, E. Celikkan, and S. Tang, "Interactive object segmentation in 3d point clouds," 05 2023, pp. 2891–2897.

[8] J. Chibane, F. Engelmann, T. A. Tran, and G. Pons-Moll, "Box2mask: Weakly supervised 3d semantic instance segmentation using bounding boxes," 2022.

[9] T. He, C. Shen, and A. van den Hengel, "Pointinst3d: Segmenting 3d instances by points," *ArXiv*, vol. abs/2204.11402, 2022.

[10] M. Zhong, X. Chen, X. Chen, G. Zeng, and Y. Wang, "Maskgroup: Hierarchical point grouping and masking for 3d instance segmentation," in *IEEE International Conference on Multimedia and Expo, ICME 2022, Taipei, Taiwan, July 18-22, 2022*. IEEE, 2022, pp. 1–6.

[11] X. Shen, Z. Zhu, L. Xie, and Y. Cui, "3d object detection, instance segmentation and classification from 3d range and 2d color images," Ph.D. dissertation, USA, 2021, aAI28262288.

[12] L. Shao, Y. Tian, and J. Bohg, "Clusternet: 3d instance segmentation in rgb-d images," 2018.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580–587.

[14] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[16] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, 2005.

[17] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[18] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *European conference on computer vision*. Springer, 2014, pp. 346–361.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016, pp. 779–788.

[21] H. Ibrahem, A. Salem, and H.-S. Kang, "Dts-net: Depth-to-space networks for fast and accurate semantic object segmentation," *Sensors*, vol. 22, 01 2022.

[22] J. Schult, F. Engelmann, A. Hermans, O. Litany, S. Tang, and B. Leibe, "Mask3d: Mask transformer for 3d semantic instance segmentation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8216–8223.

[23] Y. Jin, L. Xiangfeng, W. Yang, S. Xie, and T. Liu, "A panoramic segmentation network for point cloud," *IOP Conference Series: Earth and Environmental Science*, vol. 440, p. 032016, 03 2020.

[24] Y. Xu, S. Arai, D. Liu, F. Lin, and K. Kosuge, "Fpcc: Fast point cloud clustering-based instance segmentation for industrial bin-picking," *Neurocomputing*, vol. 494, pp. 255–268, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231222003915

[25] J. Tan, L. Chen, K. Wang, J. Li, and X. Zhang, "Saso: Joint 3d semantic-instance segmentation via multi-scale semantic association and salient point clustering optimization," *IET Computer Vision, WILEY*, vol. 15, pp. 366–379, 09 2021.

[26] Q. Lu, M. Xiao, Y. Lu, X. Yuan, and Y. Yu, "Attention-based dense point cloud reconstruction from a single image," *IEEE Access*, vol. PP, pp. 1–1, 09 2019.

[27] E. Zerman, C. Ozcinar, P. Gao, and A. Smolic, "Textured mesh vs coloured point cloud: A subjective study for volumetric video compression," 05 2020.

[28] S. Song and R. Qin, "Optimizing mesh reconstruction and texture mapping generated from a combined side-view and over-view imagery," *ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. V-2-2020, pp. 403–409, 08 2020.

[29] S. Kirchner and P. Jablonka, "Virtual archaeology: Vr based knowledge management and marketing in archaeology first results — nexts steps," in *Proceedings of the 2001 Conference on Virtual Reality, Archeology, and Cultural Heritage*, ser. VAST '01. New York, NY, USA: Association for Computing Machinery, 2001, p. 235–240.

[30] P. Reilly, ""towards a virtual archaeology."," *CAA90. Computer Applications and Quantitative Methods in Archaeology 1990 (BAR International Series 565)*, vol. Tempus Reparatum, Oxford,, pp. 403–409, (1991).

[31] M. Forte, "6.1 virtual reality, cyberarchaeology, teleimmersive archaeology," *3D Recording and Modelling in Archaeology and Cultural Heritage*, p. 115, 2014.

[32] F. Liarokapis, P. Kouřil, P. Agrafiotis, S. Demesticha, J. Chmelík, and D. Skarlatos, "3d modelling and mapping for virtual exploration of underwater archaeology assets," *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII-2/W3, pp. 425–431, 02 2017.

[33] M. Cozza, S. Isabella, P. Di Cuia, A. Cozza, R. Peluso, V. Cosentino, L. Barbieri, M. Muzzupappa, and F. Bruno, "Dive in the past: A serious game to promote the underwater cultural heritage of the mediterranean sea," *Heritage*, vol. 4, no. 4, pp. 4001–4016, 2021. [Online]. Available: https://www.mdpi.com/2571-9408/4/4/220

[34] M. Nawaf, P. Drap, M. Ben-Ellefi, E. Nocerino, B. Chemisky, T. Chassaing, A. Colpani, V. Noumossie, K. Hyttinen, J. Wood, T. Gambin, and J. C. Sourisseau, "Using virtual or augmented reality for the time-based study of complex underwater archaeological excavations," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. VIII-M-1-2021, pp. 117–124, 2021. [Online]. Available: https://isprs-annals.copernicus.org/articles/VIII-M-1-2021/117/2021/

Assist.Prof.Dr Maad Kamal Al-anni Dr. Maad Kamal Al-anni currently serves as a faculty member in the College of Engineering, Department of Computer Engineering, Al-Iraqia University. He is also involved with the World University Ranking as a mentor at Al-Iraqia University (AIU) and is the founder of the Research Center called Dynamic Casual Model and Brain Study Center at AIU, Ministry of Higher Education and Scientific Research, Iraq. He obtained his Ph.D. degree through an Indian Council for Cultural Relations (ICCR) scholarship at the University of Pune, India, from February 2005 to April 2010. Prior to this, he received a Master's degree in Computer Science from the University of Baghdad, Iraq, in October 2003, and a Bachelor's degree in Computer Science from Iraq in October 2001.

Pierre DRAP, Professor (Researcher) Dr.Pierre Drap is affiliated with Aix-Marseille University, CNRS, ENSAM, and the University of Toulon, working at LIS UMR 7020 in Marseille, France. With extensive expertise in his field, Drap's research focuses on various aspects of computer science and engineering. He has contributed significantly to the advancement of knowledge in areas such as artificial intelligence, robotics, and 3D Survey. Drap's work has been published in reputable journals and presented at international conferences. As a dedicated researcher and educator, he is committed to mentoring students and collaborating with colleagues to address complex challenges.