Proceedings of the 2025 SICE Festival with Annual Conference
September 9-12, 2025, Chiang Mai, Thailand

Applying Segment Anything Models for Analysis of Underwater Images Obtained Inside the Fukushima Daiichi Reactor

Hirokazu Madokoro[1†] and Stephanie Nix[2]

[1]Department of Software and Information Science, Iwate Prefectural University, Takizawa, Japan
(Tel: +81-19-694-2698; E-mail: hirokazu_m@iwate-pu.ac.jp)
[2]Department of Software and Information Science, Iwate Prefectural University, Takizawa, Japan
(Tel: +81-19-694-2500; E-mail: nix_s@iwate-pu.ac.jp)

Abstract: This study evaluates three advanced segmentation models of vanilla Segment Anything Model (SAM), FastSAM, and Semantic-SAM for segmentation of underwater images obtained using a remotely operated vehicle (ROV) to provide detailed and intuitive representations of the fuel debris and surrounding structures. Experimental results demonstrate that FastSAM achieves approximately 50 times faster performance compared to vanilla SAM while maintaining comparable accuracy and exceptional speed. It effectively segments diverse objects, including rectangular boxes and long bars, in images with gradually decreasing water transparency, showcasing its robust segmentation capabilities. However, the lack of an attention framework limits FastSAM's ability to distinguish between remaining regions, resulting in segmentation outputs primarily useful for predefined ranges. These limitations highlight opportunities to refine models like Semantic-SAM, enhancing their adaptability and versatility across diverse applications, including complex underwater environments.

Keywords: SAM, Segmentation, Transformer, CNN, and ROV.

1. INTRODUCTION

Massive tsunami waves, towering and relentless in the aftermath of the Great East Japan Earthquake, crashed into the Fukushima Daiichi Nuclear Power Plant (hereinafter referred to as 1F) on March 11, 2011. This devastating wall of water led to a complete loss of alternating current power at the Tokyo Electric Power Company (TEPCO)-operated facility and triggering what would soon become one of the world's most serious nuclear accidents. The first, second, and third reactors at 1F experienced meltdowns, which led to a substantial quantity of debris from the nuclear fuel and reactor components, both inside and outside the containment vessel [1]. An estimated 880 tons of debris remain within the containment vessel (PCV) [2]. The radiological levels inside are extremely high, while the physical and chemical properties of the debris display significant heterogeneity and complexity. Over the course of more than 40 years of long-term decommissioning, the retrieval of fuel debris is considered one of the most technically demanding tasks [3]. From a safety standpoint, it plays a crucial role in maintaining the integrity of the containment structure to prevent the release of radioactive materials. The retrieval of fuel debris is an unprecedented task requiring remote operations in high-radiation environments. Consequently, advanced technology and meticulous planning are essential for ensuring the safe execution of this operation.

In developing a strategy for fuel debris retrieval, it is essential to thoroughly understand the internal environment of the 1F containment vessel (PCV). Specifically, information obtained through direct visual inspections plays a crucial role in the detailed examination and precise planning required for retrieval methods. However, due to the unique challenges of operations in high-radiation environments, it is crucial to meticulously plan the work schedule and ensure operators have a thorough understanding of the working space. Careful planning and situational awareness are indispensable for safely executing these operations. To respond to these requirements, the development of technology that can comprehensively visualize the work environment in three dimensions, based on investigative data, is desirable. Due to the limitations in two-dimensional (2D) data processing and visualization, there is often insufficient spatial understanding derived from conventional 2D images, which can lead to inadequate decision-making processes in complex work environments. To significantly enhance work planning quality and reduce operational risks, the development of three-dimensional (3D) models is expected to offer a more intuitive understanding of the locations, conditions, and spatial relationships between fuel debris and surrounding structures.

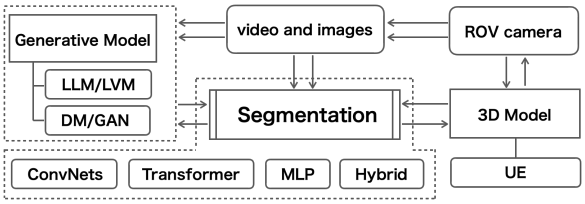

Fig. 1 Overall architecture of our system prototype.

In our research project, as reported in [4, 5], we are developing a novel approach to generate a 3D model of the work environment based on video data derived from internal inspection processes of the 1F PCV. Our method aims to provide detailed and intuitive visualizations of fuel debris and surrounding structures, enabling more accurate spatial understanding in complex operational sce-

† Hirokazu Madokoro is the presenter of this paper.

Table 1 Comparison of SAM frameworks

Feature	Vanilla SAM	FastSAM	Semantic-SAM
Architecture	ViT	CNN	ViT
Speed	Slow	Fast	Slow
Accuracy	Highest	Moderate	High
Semantic	No	No	Yes
Resource	High	Low	High
Zero-shot	High	Normal	High

narios. Fig. 1 illustrates the overall architecture of our system prototype. Our goal is to achieve precise 3D modeling of the work environment by selectively employing stereo reconstruction techniques that leverage extracted features from static or video images captured at designated time slots. Our approach aims to provide detailed and accurate visualizations of fuel debris and surrounding structures. Our approach focuses on constructing a precise 3D model of the work environment by leveraging detailed information about the positions and shapes of critical structures and obstacles extracted from video data obtained during inspections. This enables a more accurate understanding of the workspace, enhancing the quality of operation planning and reducing risks during tasks against operations. Our ultimate objective in this approach is to effectively extract spatial information from limited observational data, thereby enhancing the safety and efficiency of fuel debris retrieval operations. Moreover, we anticipate that advancements in decommissioning will be achieved by combining enhanced safety measures for workers with remote operation technologies, which enable operations in environments with severe radiation levels. In this paper, we demonstrate the potential of using semantic segmentation and instance segmentation by applying deep learning frameworks to various objects extracted from static images derived from video footage. This serves as a preliminary step toward generating 3D models for future operations aimed at enhancing safety and efficiency during fuel debris retrieval.

2. SEGMENTATION WITH THREE SAM FRAMEWORKS

The Segment Anything Model (SAM) [6] is a foundational framework for image segmentation [7]. While the vanilla SAM has been widely adopted, its variants have further expanded its applicability. These models are increasingly used in diverse tasks such as object detection, scene understanding, and 3D reconstruction, demonstrating their versatility across computer vision domains [8]. The vanilla SAM [6] utilizes a Vision Transformer (ViT) [9] backbone to encode images and employs prompt encoders to handle diverse input types, such as points, bounding boxes, and masks. While highly accurate and versatile, the vanilla SAM is computationally intensive, requiring substantial graphical processing unit (GPU) resources and processing time that limits its suitability for real-time applications. Its primary strength lies in zero-shot generalization across a wide range of visual domains [10], making it adaptable to novel tasks without retraining. However, the model performs instance segmentation only, lacking explicit semantic understanding [11].

SAM is a framework capable of segmenting any object, offering broad versatility in image analysis. Its key properties and performance characteristics have established a transformative approach in diverse applications across image recognition and segmentation. The advent of SAM has enabled detailed understanding of the shapes and positions of objects in images, facilitating precise segmentation tasks that discriminate object labels in every pixel. Although conventional segmentation models were limited to specific classes in object and stuff segmentation [12], SAM addressed this limitation by enabling fine-tuning for any objects. This enhancement significantly broadened the application range of segmentation.

One of the key properties of SAM [13] is its ability to effectively use limited labeled data, reducing reliance on extensive labeled datasets for unsupervised learning. This unsupervised training mechanism substantially lowers the cost of gathering labeled datasets. Moreover, context learning and analogical reasoning enabled SAM to rapidly adapt to novel classes without prior learning, as well as to enhance its generalization capability. Simultaneously, a novel benchmark dataset named SA-1B [13] was proposed for performance evaluation. The thorough and balanced performance evaluation provided by the SA-1B dataset establishes SAM as a foundational model. This benchmark includes a diverse range of objects and scenes not typically found in conventional segmentation benchmarks, thereby enabling more robust and representative performance assessment.

FastSAM [14] reduces the computational cost of the vanilla SAM [6] by replacing its complex ViT backbone with a lighter Convolutional Neural Network (CNN) architecture. This optimization significantly reduces inference time from seconds to milliseconds per image while preserving high-quality segmentation results. Moreover, FastSAM achieves its speed improvements through architectural optimizations and techniques like knowledge distillation, which transfer capabilities from the larger teacher model. While this approach sacrifices some accuracy compared to the vanilla SAM, it enables real-time applications on edge devices and mobile platforms, making segmentation technology accessible for time-sensitive use cases in resource-constrained environments.

Semantic-SAM [15] extends the vanilla SAM framework by incorporating semantic understanding capabilities, thereby overcoming its primary limitation [6]. By integrating semantic feature extraction modules, Semantic-SAM not only identifies object boundaries but also classifies them into semantically relevant categories. Moreover, Semantic-SAM [15] integrates zero-shot segmentation capabilities with pre-trained semantic knowledge, enabling applications that demand both precise object delineation and classification. The semantic enhancement typically incurs a modest increase in computational cost compared to the vanilla SAM but delivers substantially

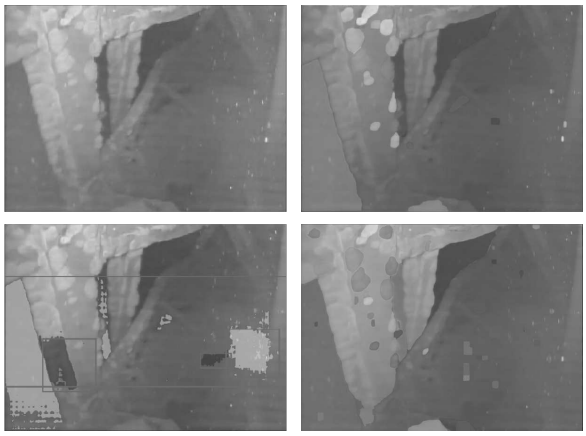

Fig. 2 Segmentation results with vanilla SAM. Upper left: input image. Upper right: instance segmentation result. Bottom left: semantic segmentation result inside BB. Bottom right: panoptic segmentation result.

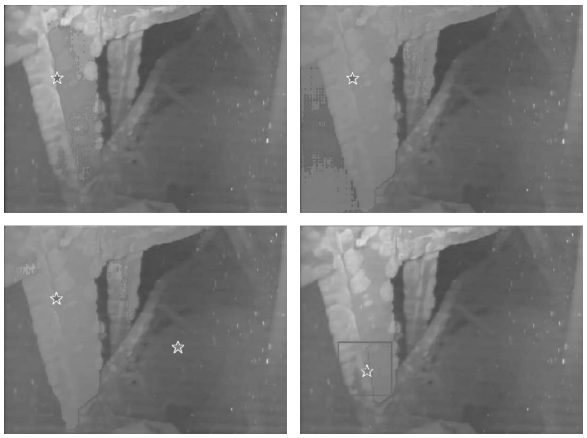

Fig. 3 Segmentation results using vanilla SAM with point-based interactions. Upper left and right: segmentation results of different masks generated from multiple points in the same positive region. Bottom left: segmentation result with both positive and negative points. Bottom right: segmentation result inside of the BB with negative point.

improved performance for tasks requiring object recognition.

As shown in Table 1, each SAM variant offers distinct tradeoffs: the vanilla SAM [6] provides exceptional accuracy and generalization at high computational cost; FastSAM [14] sacrifices some precision for dramatically improved speed and efficiency; Semantic-SAM [15] adds valuable semantic understanding while maintaining advanced segmentation quality. The choice between these models depends on specific application requirements—whether prioritizing accuracy with vanilla SAM, speed and low memory usage with FastSAM, or semantic understanding with Semantic-SAM. These models exemplify the rapid evolution of foundation models in computer vision, showcasing how a single architectural approach can be adapted to address diverse practical constraints while expanding capabilities for real-world applications. In this study, we evaluate the basic performance and functionality of the vanilla SAM, FastSAM, and Semantic-SAM for underwater images captured by a remotely operated vehicle (ROV)-mounted camera. These images are publicly available on the official TEPCO website.

3. SEGMENTATION RESULTS

3.1. Segmentation results with Vanilla SAM

Figure 2 depicts the segmentation results using the vanilla SAM. The right upper panel shows an instance segmentation result for all pixels in the entire region of the input image, as illustrated in the upper left panel. Objects at the image boundaries are excluded from the segmentation target due to the underwater environment and limitations in lighting caused by low transparency. This leads to a concentrated distribution of object instances in the upper central and left areas. We define a BB, which can be designated as a rectangle. The bottom left panel shows the result of the segmentation target within the BB. The bottom right panel displays the panoptic segmentation result, where all pixels are labeled using the vanilla SAM.

The vanilla SAM includes functions that allow users to instruct their intentions through minimal interactions, specifying positive and negative points using a mouse interface. We validated the effectiveness of segmentation using these pointing designations because it is not unrealistic to obtain ground truth (GT) annotations for each pixel of objects in states like those found after the core meltdown at the 1F. Fig. 3 depicts segmentation results for three types: using only positive points, both positive and negative points, and only negative points inside the BB. The upper left and right panels depict segmentation results obtained using the only positive point of different mask with the multiple mask option. The green markers in the image indicate the positive points provided through mouse clicks. In the left area of the image, regions with similar properties are segmented as independent instances.

The bottom-left panel shows the segmentation result using the negative points designated in the central region of the image. These points are marked by red star marks. In regions without segmentation, no distinct features are visible. As a result, the bounding box (BB) within the green-colored rectangular frames shown in the experimental results on the right panel is defined based on the partial positive regions identified in the segmentation output of the bottom-left panel. These regions are now designated as negative. The segmentation results for the left and right sides of the pillar structure differed.

3.2. Segmentation results with FastSAM

One of the challenges associated with SAM is its high computational demand. In environments where GPUs are not readily available or when immediate response is required, this demand can become a limiting factor. This

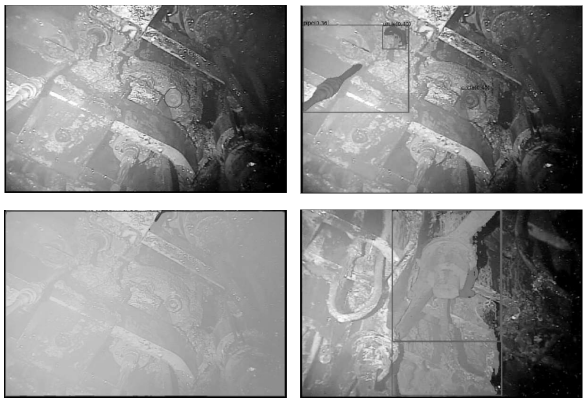

Fig. 4 Instance segmentation result using FastSAM for highly transparent images

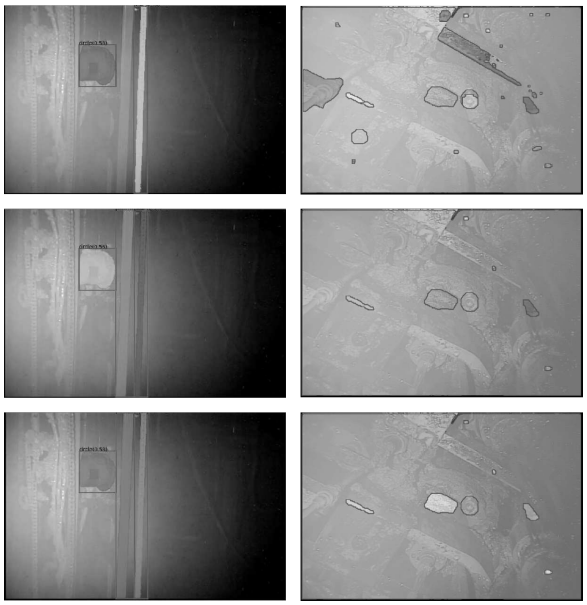

Fig. 5 Segmentation results using FastSAM in low transparency scenes. Left: segmentation for low transparency images. Right: segmentation for very low transparency images.

experiment evaluates FastSAM for real-time processing of large-scale ROV video data, addressing the need for rapid analysis. FastSAM addresses the computational burden of vanilla SAM by incorporating a Transformer backbone architecture, which decomposes the segmentation task into two steps. In the first step, it performs full instance segmentation across all objects. In the second step, it leverages prompts as instructions to refine or guide the segmentation process. Recently, You Only Look Once (YOLO) has become a well-known object detection model based on CNN backbones. Using YOLOv8-seg, a model from the YOLO family [16], FastSAM performs instance segmentation across the entire target image. Following this approach, point prompting, box prompting, and text prompting are used to select specific regions of interest. This method enables precise identification of target objects. Compared to the vanilla SAM, FastSAM demonstrates approximately 50 times faster inference speed while maintaining equivalent accuracy. Comparison experiments across multiple benchmark datasets revealed that FastSAM achieves outstanding high-speed performance while preserving comparable accuracy and efficiency relative to vanilla SAM. Furthermore, its enhanced usability makes it well-suited for real-world applications such as anomaly detection, small-object tracking, and salient object segmentation.

FastSAM performs BB extraction and instance segmentation in parallel. Fig. 4 depicts the instance segmentation results achieved using FastSAM. The upper-right and bottom-left panels show the results obtained by adjusting meta-parameters to control the granularity of the extracted instances. The bottom-right panel displays the result obtained by applying a different scene image while maintaining the same temperature settings as those in the central panel. Although FastSAM can extract highly salient regions, it is unable to partition the remaining areas into distinct instances. Consequently, the results are limited to expressing feedback for fixed ranges because FastSAM lacks an attention framework, which leads to the use of CNN-based backbones prioritizing rapid processing over precision.

We subsequently applied FastSAM to two images featuring gradually decreasing transparency in water: one with moderate transparency and the other with very low transparency. The left panels in Fig. 5 depict segmentation results for the moderately transparent image. FastSAM effectively segmented parts of both rectangular boxes and long bars, demonstrating its robust segmentation capabilities. The right panels show segmentation results for the very low transparency image, where the bounding box (BB) encompasses the entire image. While these results classify water regions as objects, it should be noted that water is typically categorized as a ”stuff” or ”substance” class in many open benchmarks. As such, FastSAM provides panoptic segmentation results [13], combining instance and semantic segmentation. In contrast, instances are extracted based on the overlap between BBs and water regions. As the temperature parameter changes, the extracted targets vary; however, the nuts located in the central and left regions remain consistently identified. In this case, the results in the left panels show that the nuts are divided into two independent regions.

3.3. Segmentation results with Semantic-SAM

Semantic-SAM [15] is a groundbreaking image segmentation model that recognizes objects and their segmented parts with high granularity. Semantic-SAM comprises two core features. The first enables the extraction of detailed semantic information, allowing it to distinguish objects and their partial classes by integrating datasets at varying levels of granularity. The second feature supports multi-granularity correspondence, which can be achieved through the combination of multiple GT masks with single-instance mask generation via interactions such as mouse clicking. This capability arises from the introduction of multi-granularity training during the

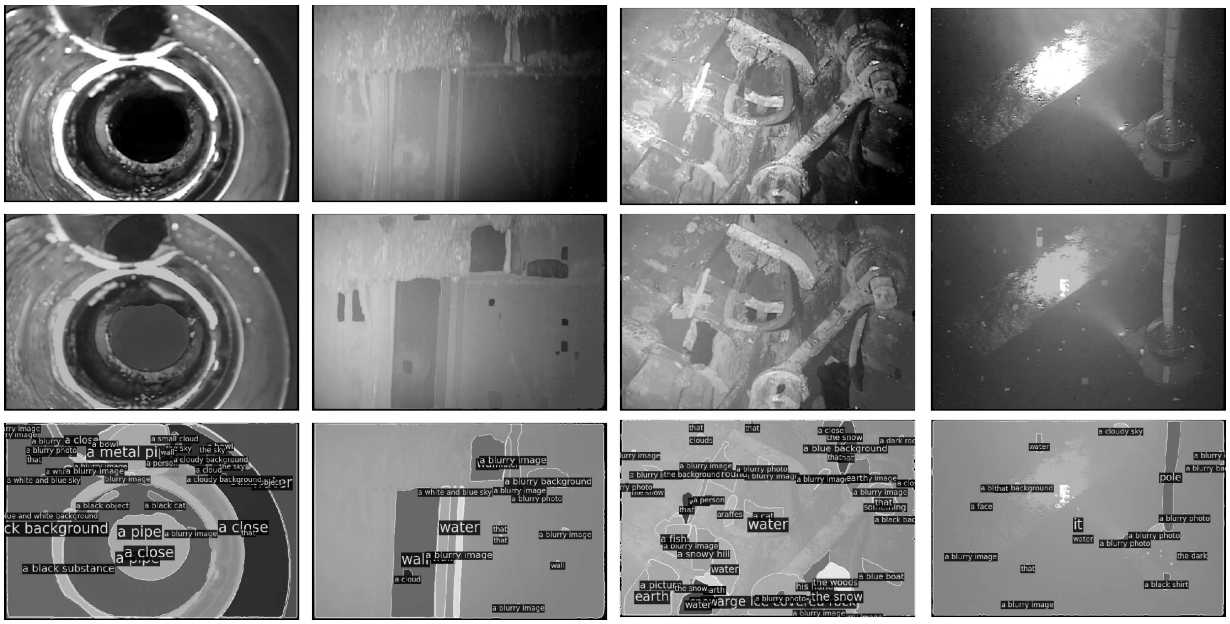

Fig. 6 Segmentation Results with Semantic-SAM. Upper: Input Images. Middle: Segmentation Results. Bottom: Labeled Segmentation Results.

pre-training phase.

Semantic-SAM was trained not only during its pre-training phase using the SA-1B datasets proposed and curated in the original SAM study but also through continual pre-training with supplementary datasets, including generic segmentation datasets and partial object segmentation datasets. This multi-step training enables Semantic-SAM to perform semantic recognition and multi-granularity classification. Furthermore, Semantic-SAM demonstrates enhanced accuracy and efficiency compared to vanilla SAM by leveraging the SA-1B dataset in conjunction with additional datasets across multiple segmentation tasks, including panoptic segmentation [13] and part segmentation [17]. Semantic-SAM addresses the limitation of fixed granularity in conventional segmentation models by offering multi-granularity segmentation capabilities. This flexibility enhances object recognition and contextual understanding across a wide range of applications.

Figure 6 displays segmentation results with object and stuff automatically assigned labels achieved using Semantic-SAM. For each image, the upper panels show the input images, the middle panels depict the segmentation outputs, and the bottom panels highlight the labeled segmentation results. The difference in instance granularity between the central and right panels arises from the omission of certain results in the right panel, particularly for tiny instances that lack distinct regions for labeling.

The limitation stems from the labeling strategy during SAM pre-training, particularly within the spatial range of the input images. We propose that this challenge can be mitigated through supervised fine-tuning (SFT) using our original image datasets collected under the 1F environment. This approach enables dynamic label reallocation and enhanced contextual learning. However, the effort required for annotation-based label adjustments, along with the computational cost of training, introduces additional challenges. As a post-processing step following SFT, the integration of the Retrieval Augmented Generation (RAG) framework may facilitate automatic label reallocation without modifying weight parameters. This is achieved through continual pre-training, which preserves model efficiency while improving adaptability to new tasks.

4. CONCLUSION

This study evaluated three advanced segmentation models of the vanilla SAM, FastSAM, and Semantic-SAM for ROV underwater images to provide detailed and intuitive representations of fuel debris and surrounding structures at 1F PCV. The experimental results demonstrated that FastSAM achieved approximately 50 times faster inference speed compared to vanilla SAM while preserving comparable accuracy and strong real-time capabilities. It effectively segments various objects, including rectangular boxes and long bars, in images with progressively reduced water opacity, showcasing its robust segmentation capabilities. However, FastSAM lacks an attention framework, which limits its capacity to accurately segment remaining areas as separate instances. Although the results were primarily focused on fixed-range scenarios, Semantic-SAM addresses the inherent limitation of fixed granularity in traditional models by enabling flexible segmentation across different granularities.

Future tasks include addressing identified limitations, such as label assignment during SAM pre-training, which can be alleviated through SFT using original image datasets. Future research should aim to optimize the trade-offs among processing speed, accuracy, and semantic understanding in image segmentation models, particularly in the context of complex and dynamic real-world

environments. This includes underwater images obtained from ROVs at 1F, which are critical for subsequent operations aimed at enhancing safety and efficiency in fuel debris retrieval, which remain a central objective in the long-term decommissioning strategy of the facility.

ACKNOWLEDGMENTS

This work was supported by the JAEA Nuclear Energy S&T and Human Resource Development Project (Grant Number: JPJA23O23813888) The authors would like to thank our lab students for running experiments for the data included in this paper..

REFERENCES

[1] Christophe Journeau, Damien Roulet, Emmanuel Porcheron, Pascal Piluso, and Christophe Chagnot and. Fukushima daiichi fuel debris simulant materials for the development of cutting and collection technologies. *Journal of Nuclear Science and Technology*, 55(9):985–995, 2018.

[2] Hirohisa Tanaka, Sogo Iwata, Tadasuke Yamamoto, Tomohito Nakayama, Shinya Uegaki, Tomoaki Kita, Atsuhiko Terada, Daiju Matsumura, Masashi Taniguchi, and Ernst-Arndt Reinecke. A proposal of hydrogen safety technology for decommissioning of the fukushima daiichi nuclear power station. *International Journal of Hydrogen Energy*, 2025.

[3] Kaiqiang Zhang, Alexandros Plianos, Luca Raimondi, Fumiaki Abe, Yoshimasa Sugawara, Ipek Caliskanelli, Alice Cryer, Justin Thomas, Salvador Pacheco-Gutierrez, Chris Hope, Ronan Kelly, Masaki Sakamoto, Tomoki Sakaue, Wataru Sato, Shu Shirai, Yolande Smith, Matthew Goodliffe, Harun Tugal, and Robert Skilton and. Towards safe, efficient long-reach manipulation in nuclear decommissioning: a case study on fuel debris retrieval at fukushima daiichi. *Journal of Nuclear Science and Technology*, 62(1):1–16, 2025.

[4] Keita Nakamura, Toshihide Hanari, Taku Matsumoto, Kuniaki Kawabata, and Hiroshi Yashiro. A study on the effects of photogrammetry by the camera angle of view using computer simulation. *Journal of Robotics and Mechatronics*, 36(1):115–124, 2024.

[5] Toshihide Hanari, Keita Nakamura, Takashi Imabuchi, and Kuniaki Kawabata. Image selection method from image sequence to improve computational efficiency of 3d reconstruction: Application of fixed threshold to remove redundant images. *Journal of Robotics and Mechatronics*, 36(6):1537–1549, 2024.

[6] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, October 2023.

[7] Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, Hao Peng, Jianxin Li, Jia Wu, Ziwei Liu, Pengtao Xie, Caiming Xiong, Jian Pei, Philip S. Yu, and Lichao Sun. A comprehensive survey on pretrained foundation models: a history from bert to chatgpt. *International Journal of Machine Learning and Cybernetics*, November 2024.

[8] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3523–3542, 2022.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[10] Udit Maniyar, Joseph K J, Aniket Anand Deshmukh, Urun Dogan, and Vineeth N Balasubramanian. Zero shot domain generalization, 2020.

[11] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. A comprehensive survey on segment anything model for vision and beyond, 2023.

[12] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2019.

[13] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[14] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything, 2023.

[15] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. Semantic-sam: Segment and recognize anything at any granularity, 2023.

[16] Juan Terven, Diana-Margarita Córdova-Esparza, and Julio-Alejandro Romero-González. A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Machine Learning and Knowledge Extraction*, 5(4):1680–1716, 2023.

[17] Tai-Yu Pan, Qing Liu, Wei-Lun Chao, and Brian Price. Towards open-world segmentation of parts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15392–15401, June 2023.