Balancing Speed and Accuracy: A Comparative Analysis of Segment Anything-Based Models for Robotic Indoor Semantic Mapping

Bruno Georgevich Ferreira[1,2,3,4][a], Armando Jorge Sousa[1,2][b] and Luis Paulo Reis[1,3][c]
[1]*FEUP - Faculty of Engineering of the University of Porto, Porto, Portugal*
[2]*INESC TEC - INESC Technology and Science, Porto, Portugal*
[3]*LIACC - Artificial Intelligence and Computer Science Laboratory, Porto, Portugal*
[4]*Edge Innovation Center, Federal University of Alagoas, Maceió, Brazil*

Keywords: Semantic Segmentation, Robotic Indoor Mapping, Semantic Mapping, Segment Anything Model (SAM), Performance Evaluation, Computer Vision, FastSAM, SAM2, MobileSAM.

Abstract: Semantic segmentation is a relevant process for creating the rich semantic maps required for indoor navigation by autonomous robots. While foundation models like Segment Anything Model (SAM) have significantly advanced the field by enabling object segmentation without prior references, selecting an efficient variant for real-time robotics applications remains a challenge due to the trade-off between performance and accuracy. This paper evaluates three such variants — FastSAM, MobileSAM, and SAM 2 — comparing their speed and accuracy to determine their suitability for semantic mapping tasks. The models were assessed within the Robot@VirtualHome dataset across 30 distinct scenes, with performance quantified using Frames Per Second (FPS), Precision, Recall, and an Over-Segmentation metric, which quantifies the fragmentation of an object into multiple masks, preventing high quality semantic segmentation. The results reveal distinct performance profiles: FastSAM achieves the highest speed but exhibits poor precision and significant mask fragmentation. Conversely, SAM 2 provides the highest precision but is computationally intensive for real-time applications. MobileSAM emerges as the most balanced model, delivering high recall, good precision, and viable processing speed, with minimal over-segmentation. We conclude that MobileSAM offers the most effective trade-off between segmentation quality and efficiency, making it a good candidate for indoor semantic mapping in robotics.

1 INTRODUCTION

Semantic segmentation is the process of labelling each pixel in an image with a class. It is a fundamental problem in computer vision that provides a detailed, pixel-level understanding of a scene (Guo et al., 2018), which is critical for a wide range of applications, from medical image analysis (Syam et al., 2023; Osei et al., 2023) and autonomous driving (Lian et al., 2025) to agriculture (Luo et al., 2024), remote sensing (Huang et al., 2024), and infrastructure inspection (Wang et al., 2022). In the domain of robotics and autonomous systems, semantic segmentation is particularly crucial as it enables the creation of rich, machine-readable representations of the environment, commonly known as semantic maps (Liu et al., 2025). These maps enhance a robot's ability to navigate, interact with its surroundings, and perform complex tasks.

The field has seen rapid evolution, moving from traditional methods to deep learning architectures, which have become the state-of-the-art (Guo et al., 2018; Sohail et al., 2022). The landscape of deep learning models for semantic segmentation is diverse, encompassing architectures based on Convolutional Neural Networks (CNNs) and more recently, Vision Transformers (ViTs) and their variants (Sohail et al., 2022; Thisanke et al., 2023). The choice of a specific model involves a critical trade-off between detection quality, computational efficiency, and suitability for a given application domain (Broni-Bediako et al., 2023). While numerous studies have compared these models, they often focus on specific domains such as remote sensing (Dahal et al., 2025) or medical imag-

[a] https://orcid.org/0000-0003-1345-5103
[b] https://orcid.org/0000-0002-0317-4714
[c] https://orcid.org/0000-0002-4709-1718

Ferreira, B. G., Sousa, A. J. and Reis, L. P.
Balancing Speed and Accuracy: A Comparative Analysis of Segment Anything-Based Models for Robotic Indoor Semantic Mapping.
DOI: 10.5220/0013778500003982

In *Proceedings of the 22nd International Conference on Informatics in Control, Automation and Robotics (ICINCO 2025) - Volume 1*, pages 321-328
ISBN: 978-989-758-770-2; ISSN: 2184-2809

ing (Pak et al., 2024). A comparative analysis within a controlled, simulated robotic environment is essential for evaluating model performance for indoor semantic mapping tasks.

Recent advancements have been marked by the rise of important foundation models, most notably the Segment Anything Model (SAM), which offers remarkable zero-shot generalization (Kirillov et al., 2023; Zhao et al., 2023). These models, pretrained on massive datasets, can segment objects and scenes without task-specific training, presenting a new approach. Modern approaches now focus on leveraging these capabilities for semantic mapping by adapting them for specific domains (Cheng et al., 2025), integrating them into multimodel systems for real-time tracking in dynamic environments (Khajarian et al., 2025), and using them to power semi-automated annotation pipelines that dramatically accelerate the creation of high-quality datasets (He et al., 2025; Lian et al., 2025). Efforts are also underway to make these large models more efficient for real-world deployment on edge devices, leading to lightweight variants like FastSAM and MobileSAM (Zhao et al., 2023; Zhang et al., 2023), and to extend their capabilities to new modalities like video with the introduction of SAM 2 (Ravi et al., 2024).

For instance, while ViTs may excel in some domains, CNNs have been shown to be more effective or efficient in others, such as in specific surgical tasks (Pak et al., 2024). Similarly, in the context of medical imaging, CNNs have also demonstrated superior performance (Osei et al., 2023). This highlights an interesting gap, where the performance of a given model is deeply related with the specific characteristics of the application domain and by the dataset used for evaluation. Therefore, direct empirical evaluation is necessary to determine the most suitable model for a specific use case.

This paper addresses this gap by presenting a comparative analysis of the performance and detection quality of three distinct semantic segmentation models: FastSAM (Zhao et al., 2023), MobileSAM (Zhang et al., 2023), and SAM 2 (Ravi et al., 2024). The evaluation is conducted specifically within the context of semantic mapping for robotics, using the Robot@VirtualHome (Fernandez-Chaves et al., 2022) database to ensure the findings are relevant for indoor autonomous systems. Our contribution is to provide a clear comparison between these models that can guide the selection of semantic segmentation models in the context of semantic mapping.

2 RELATED WORKS

The advent of deep learning has driven rapid evolution in the field of semantic segmentation. This section reviews the resulting literature, focusing on architectural paradigms, the rise of foundation models, comparative studies, advanced methodologies, and the application of segmentation to extract relevant information to create semantic maps.

2.1 Architectural Paradigms in Semantic Segmentation

The progression of segmentation models began with methods like Fully Convolutional Networks (FCNs) and weakly supervised approaches (Guo et al., 2018; Mo et al., 2022). For a significant period, CNNs were predominant. Encoder-decoder architectures, such as U-Net, excelled at generating dense, pixel-wise predictions while preserving spatial information (Zhang et al., 2021). Other prominent CNNs, including PSPNet and DeepLabV3+, utilize spatial pyramid pooling to aggregate multi-scale context (Sohail et al., 2022), with systematic comparisons evaluating their performance across various domains (Wang et al., 2022). More recently, ViTs like SegFormer, Swin Transformer, and Mask2Former have emerged as powerful alternatives (Thisanke et al., 2023). The self-attention mechanism in ViTs allows for more effective modeling of long-range spatial dependencies compared to the localized receptive fields of CNNs, a development closely tracked in fields such as remote sensing (Huang et al., 2024).

2.2 Comparative Analyses: CNNs versus Transformers

The advent of ViTs prompted direct performance comparisons with CNNs. Transformer-based models often demonstrate superior accuracy, achieving higher mean Intersection over Union (mIoU) scores in applications like remote sensing (Dahal et al., 2025), medical imaging (Syam et al., 2023), watershed classification (He et al., 2025), and natural disaster assessment (Asad et al., 2023). However, this superiority is not universal. In specific contexts, such as 3D brain tumor segmentation (Osei et al., 2023) and complex robotic surgery environments (Pak et al., 2024), established CNNs have been shown to outperform Transformers. These conflicting findings underscore that architectural performance is highly dependent on the domain, dataset, and task. A frequently noted trade-off is the superior computational efficiency (e.g., FLOPS, inference time) of CNNs for

comparable accuracy, highlighting that ViTs' accuracy gains can come at a significant computational cost (Dahal et al., 2025; Broni-Bediako et al., 2023).

2.3 Foundation Models and Advanced Methodologies

A new paradigm has emerged with large-scale, pretrained foundation models like the SAM, known for its zero-shot segmentation capabilities. Current research focuses on adapting these computationally intensive models for practical use. This includes creating lightweight versions like FastSAM and MobileSAM for edge devices (Zhao et al., 2023; Zhang et al., 2023), and extending capabilities to video with models like SAM 2 (Ravi et al., 2024). These models are being integrated into advanced frameworks for semi-automated annotation (He et al., 2025; Lian et al., 2025), semi-supervised learning in data-scarce domains via pseudo-labeling (Xu et al., 2025), and specialized tasks such as remote sensing, using automated prompt generation (Cheng et al., 2025). For real-time applications like AR-guided surgery, multi-model systems are being developed that combine foundation models with other specialized networks to balance accuracy and speed (Khajarian et al., 2025). Other advanced methods include creating hybrid CNN-Transformer architectures (e.g., LETNet) to leverage the strengths of both (Xu et al., 2022). Concurrently, there is a growing emphasis on fair and holistic benchmarking, moving beyond isolated metrics to evaluate the entire pipeline's trade-offs between quality (mIoU) and efficiency (FPS, power consumption) for real-world deployment (Lee et al., 2022; Mo et al., 2022; Broni-Bediako et al., 2023).

2.4 Application in Semantic Mapping

In robotics, semantic segmentation is fundamental for constructing detailed environmental models. For instance, systems now use models like YOLOv8s-seg to build object-level semantic maps of forests for SLAM (Liu et al., 2025), or create vectorized maps of traffic signs for low-cost autonomous vehicle localization (Lian et al., 2025). The development of realistic simulators with ground-truth data, such as Robot@VirtualHome, is critical for training and validating these systems before deployment (Fernandez-Chaves et al., 2022). This highlights that the utility of the final semantic map is directly dependent on the accuracy and efficiency of the underlying segmentation model, positioning our work to evaluate models for this specific application.

3 METHODOLOGY

This section describes the experimental methodology used to assess the performance of three semantic segmentation models — FastSAM, MobileSAM, and SAM 2 — within the Robot@VirtualHome dataset's simulated real-world environment. The methodology includes the experimental setup, dataset details, segmentation model specifications, evaluation metrics, and hardware specifications.

3.1 Experimental Setup

To assess the efficacy of three semantic segmentation models—SAM2, FastSAM, and MobileSAM—in realistic scenarios, they were first adapted to operate within the Robot@VirtualHome simulation environment. Subsequently, the models were executed across all 30 distinct homes available in the simulator. The predefined "Wandering" trajectory was selected for model validation because it provides a sufficient and representative path for comprehensive analysis. This trajectory initializes a robot at a random location, which then navigates the environment while collecting data.

3.2 Dataset

The study utilizes the Robot@VirtualHome dataset, which features 30 simulated homes designed from real-world layouts to ensure realistic object placements. For each of the approximately 16,400 data points, the dataset provides an RGB image, a depth image, and a ground truth segmentation mask. The RGB images serve as the input for the models, with the provided masks used for evaluation. Data on the robot's pose and from a simulated LiDAR sensor are also included.

3.3 Semantic Segmentation Models

The research focused on efficient models to balance performance and computational cost. The selected models include **FastSAM S** (based on YOLOv8s, 68M parameters) using half-precision, **MobileSAM** (9.66M parameters), and **SAM2** (Hiera Tiny v2.1, 38.9M parameters).

3.4 Evaluation Metrics

Model performance was assessed using four metrics: Precision, Recall, Frames Per Second (FPS), and Over-Segmentation.

3.4.1 Precision

Measures the accuracy of the segmentation, penalizing false positives and fragmented masks. A higher score is better.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3.4.2 Recall

Assesses the model's ability to identify all ground truth objects in a scene. A higher score is better.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3.4.3 Frames per Second (FPS)

Quantifies processing speed across the entire pipeline. A higher value indicates better efficiency.

$$\text{FPS} = \frac{\text{Total Frames}}{\sum T_{\text{processing}}}$$

3.4.4 Over-Segmentation

Evaluates mask fragmentation by comparing the number of predicted masks to the number of ground truth masks. A lower value is desirable.

$$\text{Over-Segmentation} = \frac{\text{Nº Pred. Masks}}{\text{Nº GT Masks}}$$

3.5 Segmentation Mapping Process

To compute the evaluation metrics, a greedy algorithm was implemented to establish a one-to-one correspondence between predicted and ground truth masks. The matching process, illustrated in Figure 1, is based on spatial proximity and overlap.

The procedure is as follows:

1. **Centroid Calculation:** The geometric centroid of each ground truth and predicted mask is computed to serve as a spatial reference point.
2. **Distance Threshold Definition:** A maximum search radius ($d_{\max}$) is dynamically set to 30% of the smaller image dimension (height or width) to filter potential matches based on centroid proximity.
3. **Greedy Mapping Algorithm:** The core of the process is an algorithm that iterates through each ground truth mask to find its best match among the available (unmapped) predicted masks. For each ground truth mask, the following criteria are applied:
   (a) *Distance Filter:* Only predicted masks whose centroids are within the $d_{\max}$ are considered candidates.
   (b) *Intersection over Union (IoU) Filter:* For the candidates that pass the distance filter, the IoU is calculated. Only predictions with an IoU greater than 10% are considered, ensuring a minimal degree of actual overlap.
   (c) *Combined Score Selection:* For candidates that satisfy both filters, a hybrid score is computed to determine the best match. The score is defined as:

   $$\text{combined\_score} = \text{IoU} - \left(\frac{d}{d_{\max}}\right) \times IoU_{\min}$$

   where d is the distance between the centroids, $d_{\max}$ is the maximum distance threshold and $IoU_{\min}$ is the minimum overlap necessary to be considered a candidate. The predicted mask with the highest `combined_score` is selected as the best match.

   This approach is considered "greedy" as it makes the locally optimal choice for each ground truth mask without guaranteeing a globally optimal set of pairings.
4. **Final Mapping Generation:** The output of this process is a map that links the identifier of each ground truth mask to the identifier of its corresponding predicted mask. This ensures a one-to-one mapping, where each predicted mask can be assigned to at most one ground truth mask. Ground truth masks with no valid match are classified as FN, while predicted masks left unmapped are classified as FP.

3.6 Hardware Specifications

All experiments were conducted on a dedicated system to ensure consistency and reproducibility. The hardware configuration consisted of an Intel Core i7-13700K CPU, an NVIDIA GeForce RTX 4090 GPU, and 64 GB of DDR5 RAM (5200 MT/s).

4 RESULTS AND DISCUSSION

The performance of the FastSAM, MobileSAM, and SAM 2 models was benchmarked using the Robot@VirtualHome dataset, focusing on relevant metrics for real-time robotic applications. This section presents a detailed analysis of the results, examining each metric in a dedicated subsection to understand the specific strengths and weaknesses of each

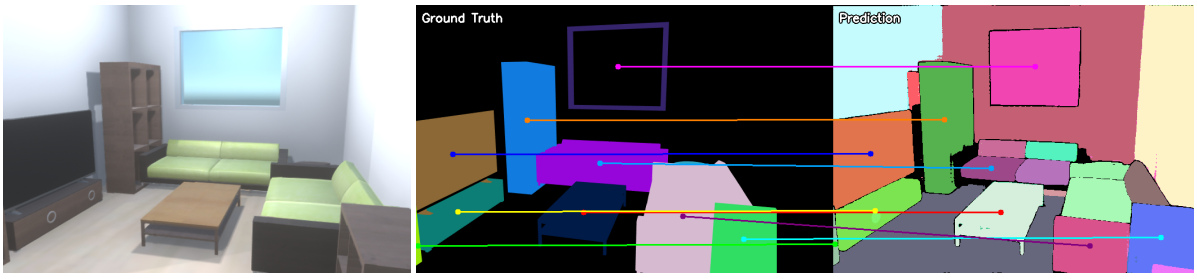

Figure 1: Comparison of ground truth segmentation with masks predicted by the SAM2 model for the Home04 scene from the Robot@VirtualHome environment. The figure displays the original input image (left), the ground truth segmentation masks (center), and the predicted masks (right). Colored lines connect the centroids of ground truth objects to their corresponding predicted segments. The results demonstrate a notable instance of over-segmentation, where the model divides single ground truth objects, such as the sofa and the wall, into multiple distinct segments.

model. The findings reveal critical trade-offs between processing speed, detection accuracy, and mask quality.

4.1 Processing Speed (FPS)

Processing speed, measured in FPS, is paramount for any model intended for real-time deployment on a robotic platform. As shown in Figure 2, there are vast differences in computational efficiency among the evaluated models.

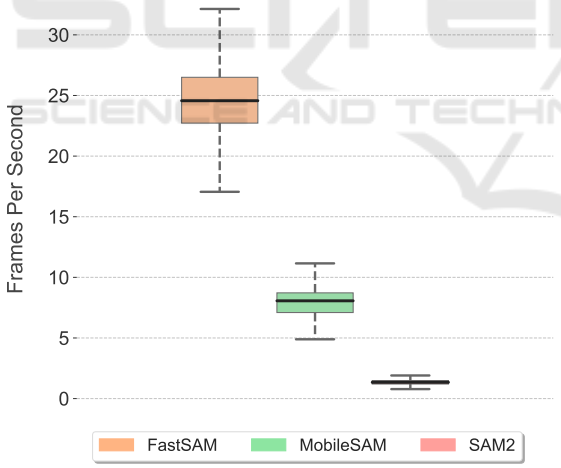

Figure 2: The chart illustrates the computational performance of three models, measured in FPS. There are significant differences in processing speed among the models. FastSAM demonstrates the highest throughput, while MobileSAM shows a more moderate performance. In contrast, SAM2 is considerably slower, making it less practical for real-time applications.

FastSAM demonstrates exceptional performance, achieving a median speed of approximately 25 FPS, establishing it as the most computationally efficient model by a large margin. This high throughput can be partially attributed to its architecture and support for half-precision processing. Following it is **MobileSAM**, with a more moderate but still viable median speed of around 8-9 FPS. At the opposite end of the spectrum, **SAM 2** is exceedingly slow, with a median performance of only 1.5 FPS. This result renders SAM 2 impractical for applications requiring rapid perception of dynamic environments, while FastSAM's speed makes it an attractive candidate from a purely computational standpoint.

4.2 Precision

Precision evaluates a model's ability to generate accurate segmentations without producing false positives. This metric is critical for creating a reliable semantic map, as low precision introduces non-existent or incorrectly classified objects into the robot's world representation. The results are presented in Figure 3.

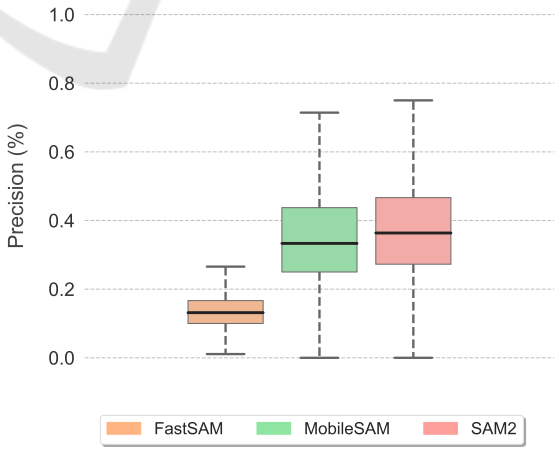

Figure 3: The box plots illustrate the distribution of precision scores, where a higher value indicates a lower rate of false positive segmentations. SAM2 achieves the highest median precision, followed closely by MobileSAM. In contrast, FastSAM shows significantly lower performance, indicating consistently low precision.

SAM 2 stands out with the highest median precision (≈ 0.37), indicating its predictions are the most reliable among the three. MobileSAM follows with respectable precision, showing a median in the 0.32–0.33 range. This suggests it produces relatively clean segmentations with a low rate of false positives and over-segmentation. In the opposite direction, FastSAM exhibits a poor precision, with a median score of just 0.13. This low score is a direct consequence of its tendency to segment large, unannotated areas of the scene, such as walls and floors, and by frequently over-segment the objects which are counted as false positives according to our evaluation protocol.

4.3 Recall

Recall, or sensitivity, measures a model's ability to segment all the ground truth objects present in a scene. High recall is essential to ensure the semantic map is complete and does not miss important environmental features. However, this metric does not account for noise and false positives. The performance of the models on this metric is shown in Figure 4.

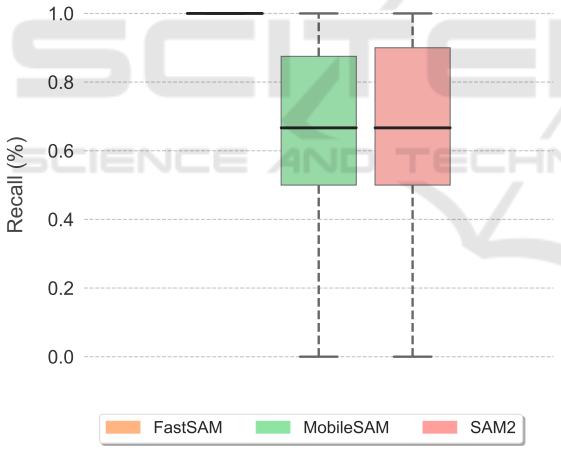

Figure 4: Recall measures a model's ability to identify all ground truth objects in an image. MobileSAM and SAM2 show similar performance. The wide range of scores for these two models indicates variability in their performance across different scenes. FastSAM achieves the highest median recall, but this high score is a consequence of the model's tendency to over-segment, detecting all true objects but also generating a high number of false positives.

MobileSAM and SAM 2 achieved a similar performance, sharing a median recall of approximately 0.67. This indicates a good capability to segment the majority of objects in a given frame, which is important for building a comprehensive map. Interestingly, the FastSAM box plot shows a median recall of 1.0. This seemingly perfect score is not an indicator of superior performance but rather an artifact of its operating principle; by segmenting almost everything in the image indiscriminately, it invariably covers all ground truth objects, but at the direct and severe cost of the low precision discussed previously.

4.4 Over-Segmentation Analysis

The over-segmentation ratio quantifies the average number of fragments the segmentation model splits a single real-world object mask into multiple predicted masks. A lower ratio is highly desirable, as high fragmentation complicates the process of creating a coherent and object-centric semantic map, requiring significant post-processing to merge segments. The results are detailed in Figure 5.

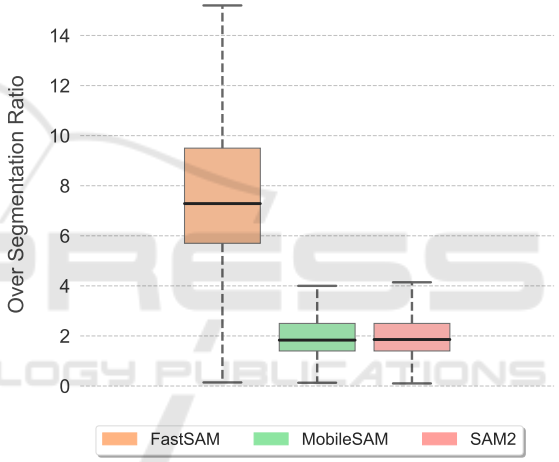

Figure 5: The over-segmentation ratio measures the average number of predicted segments generated for a single ground-truth object, where lower values indicate better performance. FastSAM shows a significantly higher median over-segmentation ratio, indicating that it tends to excessively fragment single objects into multiple segments. In contrast, MobileSAM and SAM2 demonstrate a more stable performance. This suggests that MobileSAM and SAM2 produce more coherent and usable segmentations with minimal fragmentation.

This metric reveals the most dramatic difference between the models. FastSAM is a severe outlier, with a median over-segmentation ratio of over 7. This implies that for every object in the scene, FastSAM generates, on average, more than seven distinct segments, making its output almost unusable for direct semantic mapping without complex post-processing. In contrast, MobileSAM and SAM 2 demonstrate well-controlled performance, with median ratios clustered around a much more manageable value of 2. This indicates a consistent and predictable behavior

where they occasionally split an object into two segments — a far more tractable issue for any downstream processing pipeline.

In summary, the detailed analysis of each metric reveals a clear trade-off. FastSAM, while extremely fast, suffers from quality issues, particularly in precision and over-segmentation coherence. SAM 2, while highly precise, is slow for real-time use. MobileSAM consistently performs at or near the top for quality metrics (Precision and Recall) while maintaining a moderate speed and a low over-segmentation ratio, positioning it as the most balanced and pragmatic choice for the target application.

5 CONCLUSION

This paper presented a comparative analysis of three modern, efficient variants of the SAM — specifically FastSAM (initial reference dated 2023), MobileSAM (initial reference dated 2021), and SAM 2 (initial reference dated 2024) — for the task of indoor semantic mapping. By leveraging the realistic Robot@VirtualHome simulation environment, we evaluated the models on key metrics of speed (FPS), accuracy (Precision and Recall), and mask quality (Over-Segmentation Ratio) to determine their suitability for deployment on autonomous systems.

Our findings reveal a distinct trade-off profile for each model. FastSAM operates at a remarkable speed, making it attractive for applications with stringent latency requirements. However, this velocity comes at the cost of a low precision and a high degree of mask fragmentation, rendering it unsuitable for tasks that demand accurate and coherent object identification. In direct contrast, SAM 2 delivers the highest precision and a good recall, producing high-fidelity segmentations that would be ideal for offline map creation. Unfortunately, its low frame rate makes it impractical for real-time scenarios where the environment is dynamic, or the robot is in motion.

The most compelling performance for the target application of real-time indoor semantic mapping was delivered by MobileSAM. It strikes an effective and pragmatic balance, providing high recall and good precision while maintaining a processing speed that is viable for on-the-fly robotic operations. Its low and predictable over-segmentation ratio further solidifies its position as a practical and reliable choice for generating semantic maps that are both comprehensive and accurate.

[1]https://github.com/CASIA-IVA-Lab/FastSAM

[2]https://github.com/ChaoningZhang/MobileSAM

[3]https://github.com/facebookresearch/sam2

This work contributes a clear, data-driven guide for researchers, engineers and practitioners in selecting a segmentation model for robotics. We conclude that for indoor semantic mapping, where both the accuracy of the environmental representation and the timeliness of its updates are crucial, MobileSAM currently offers the best trade-off between detection quality and computational efficiency.

Future work should aim to validate these findings on a physical robotic platform to account for real-world complexities not present in simulation. Further research could also explore the development of post-processing algorithms to merge over-segmented regions, potentially improving the usability of faster models like TinySAM. Finally, an analysis incorporating power consumption would provide an even more comprehensive understanding of model efficiency for deployment on battery-powered mobile robots.

ACKNOWLEDGEMENTS

This work was partially financially supported by: Base Funding - UIDB/00027/2020 of the Artificial Intelligence and Computer Science Laboratory – LIACC - funded by national funds through the FCT/MCTES (PIDDAC). This work is partially financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020.

REFERENCES

Asad, M. H., Asim, M. M., Awan, M. N. M., and Yousaf, M. H. (2023). Natural Disaster Damage Assessment using Semantic Segmentation of UAV Imagery. *2023 International Conference on Robotics and Automation in Industry, ICRAI 2023*.

Broni-Bediako, C., Xia, J., and Yokoya, N. (2023). Real-Time Semantic Segmentation: A brief survey and comparative study in remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 11:94–124.

Cheng, G., Zhao, Q., Lyu, S., Rizeei, H. M., Zhang, J., Li, Y., Yang, X., Jiang, R., and Zhang, L. (2025). RSAM-Seg: A SAM-Based Model with Prior Knowledge Integration for Remote Sensing Image Semantic Segmentation. *Remote Sensing 2025, Vol. 17, Page 590*, 17:590.

Dahal, A., Murad, S. A., and Rahimi, N. (2025). Heuristical Comparison of Vision Transformers Against Convolutional Neural Networks for Semantic Segmentation on Remote Sensing Imagery. *IEEE Sensors Journal*, 25:17364–17373.

Fernandez-Chaves, D., Ruiz-Sarmiento, J. R., Jaenal, A., Petkov, N., and Gonzalez-Jimenez, J. (2022). RobotVirtualHome, an ecosystem of virtual environments and tools for realistic indoor robotic simulation. *Expert Systems with Applications*, 208:117970.

Guo, Y., Liu, Y., Georgiou, T., and Lew, M. S. (2018). A review of semantic segmentation using deep neural networks. *International Journal of Multimedia Information Retrieval*, 7:87–93.

He, P., Shen, T., Wang, Y., Zhu, D., Hu, Q., Li, H., Yin, W., Zhang, Y., and Yang, A. (2025). A study on automatic annotation methods for watershed environmental elements based on semantic segmentation models. *European Journal of Remote Sensing*, 58.

Huang, L., Jiang, B., Lv, S., Liu, Y., and Fu, Y. (2024). Deep-Learning-Based Semantic Segmentation of Remote Sensing Images: A Survey. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:8370–8396.

Khajarian, S., Schwimmbeck, M., Holzapfel, K., Schmidt, J., Auer, C., Remmele, S., and Amft, O. (2025). Automated multimodel segmentation and tracking for AR-guided open liver surgery using scene-aware self-prompting. *International Journal of Computer Assisted Radiology and Surgery* .

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W. Y., Dollár, P., and Girshick, R. (2023). Segment Anything. *Proceedings of the IEEE International Conference on Computer Vision*, pages 3992–4003.

Lee, M. S., Kim, M., and Jeong, C. Y. (2022). Real-time semantic segmentation on edge devices: A performance comparison of segmentation models. *International Conference on ICT Convergence*, 2022-October:383–388.

Lian, J., Chen, S., Guo, G., Sui, D., Zhao, J., and Li, L. (2025). Lightweight semantic visual mapping and localization based on ground traffic signs. *Displays*, 90.

Liu, H., Xu, G., Liu, B., Li, Y., Yang, S., Tang, J., Pan, K., and Xing, Y. (2025). A real time LiDAR-Visual-Inertial object level semantic SLAM for forest environments. *ISPRS Journal of Photogrammetry and Remote Sensing*, 219:71–90.

Luo, Z., Yang, W., Yuan, Y., Gou, R., and Li, X. (2024). Semantic segmentation of agricultural images: A survey. *Information Processing in Agriculture*, 11:172–186.

Mo, Y., Wu, Y., Yang, X., Liu, F., and Liao, Y. (2022). Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing*, 493:626–646.

Osei, I., Appiah-Kubi, B., Frimpong, B. K., Hayfron-Acquah, J. B., Owusu-Agyemang, K., Ofori-Addo, S., Turkson, R. E., and Mawuli, C. B. (2023). Multimodal Brain Tumor Segmentation Using Transformer and UNET. *2023 20th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2023*.

Pak, S., Park, S. G., Park, J., Choi, H. R., Lee, J. H., Lee, W., Cho, S. T., Lee, Y. G., and Ahn, H. (2024). Application of deep learning for semantic segmentation in robotic prostatectomy: Comparison of convolutional neural networks and visual transformers. *Investigative and clinical urology*, 65:551–558.

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., Feichtenhofer, C., and Fair, M. (2024). SAM 2: Segment Anything in Images and Videos.

Sohail, A., Nawaz, N. A., Shah, A. A., Rasheed, S., Ilyas, S., and Ehsan, M. K. (2022). A Systematic Literature Review on Machine Learning and Deep Learning Methods for Semantic Segmentation. *IEEE Access*, 10:134557–134570.

Syam, R. F. K., Rachmawati, E., and Sulistiyo, M. D. (2023). Whole-Body Bone Scan Segmentation Using SegFormer. *2023 10th International Conference on Information Technology, Computer, and Electrical Engineering, ICITACEE 2023*, pages 419–424.

Thisanke, H., Deshan, C., Chamith, K., Seneviratne, S., Vidanaarachchi, R., and Herath, D. (2023). Semantic segmentation using Vision Transformers: A survey. *Engineering Applications of Artificial Intelligence*, 126.

Wang, J. J., Liu, Y. F., Nie, X., and Mo, Y. L. (2022). Deep convolutional neural networks for semantic segmentation of cracks. *Structural Control and Health Monitoring*, 29.

Xu, G., Qian, X., Shao, H. C., Luo, J., Lu, W., and Zhang, Y. (2025). A segment anything model-guided and match-based semi-supervised segmentation framework for medical imaging. *Medical Physics*.

Xu, Z., Guan, H., Kang, J., Lei, X., Ma, L., Yu, Y., Chen, Y., and Li, J. (2022). Pavement crack detection from CCD images with a locally enhanced transformer network. *International Journal of Applied Earth Observation and Geoinformation*, 110.

Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S.-H., Lee, S., and Hong, C. S. (2023). Faster Segment Anything: Towards Lightweight SAM for Mobile Applications.

Zhang, W., Tang, P., and Zhao, L. (2021). Fast and accurate land cover classification on medium resolution remote sensing images using segmentation models. *International Journal of Remote Sensing*, 42:3277–3301.

Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., and Wang, J. (2023). Fast Segment Anything.