https://doi.org/10.31891/2307-5732-2026-361-17
УДК 004.[93+4]

БОРКІВСЬКИЙ БОГДАН
Національний університет "Львівська Політехніка"
https://orcid.org/0000-0003-3301-476X
e-mail: bohdan.p.borkivskyi@lpnu.ua
ТЕСЛЮК ВАСИЛЬ
Національний університет "Львівська Політехніка"
https://orcid.org/0000-0002-5974-9310
e-mail: vasyl.m.teslyuk@lpnu.ua

ПОКРАЩЕННЯ РОЗПІЗНАВАННЯ ПЕРЕШКОД В ЗАКРИТИХ СЕРЕДОВИЩАХ ДЛЯ РОБОТИЗОВАНИХ СИСТЕМ

Точне розпізнавання перешкод відіграє вирішальну роль у забезпеченні безпечної та ефективної роботи робототехнічних систем у закритих приміщеннях, особливо в автомобільних додатках, де безпека та надійність мають вирішальне значення. Це дослідження зосереджується на покращенні ефективності сегментації типових перешкод у закритих приміщеннях, включаючи меблі та елементи конструкції. У дослідженні представлено огляд та аналіз останніх досягнень у галузі базових моделей сегментації, з акцентом на їхньому потенціалі в робототехнічних системах та використанні архітектур, таких як SAM2. Для оцінки їх застосовності ми адаптуємо універсальну модель сегментації, налаштовуючи її на RGB-зображення з набору даних NYU Depth V2, який містить широкий спектр внутрішніх сцен, знятих з точок зору, подібних до тих, що зустрічаються в робототехнічних системах. Інформація про глибину була виключена, щоб відобразити обмеження монокулярних камер, типових для вбудованої робототехніки. Адаптована модель досягає покращеної точності сегментації, досягаючи 82,4% mIoU, і демонструє більшу надійність у розпізнаванні перешкод у захаращених і візуально складних сценах. Результати експериментів показують постійне покращення порівняно з базовою моделлю, з приростом продуктивності на 7,6%, що підтверджує цінність адаптації для конкретної галузі для покращення сприйняття роботами. Хоча якість сегментації варіюється залежно від кількості точкових підказок (продуктивність гірша з 5 точками, ніж з 15), покращення порівняно з базовою моделлю у випадку малих вхідних даних є значно вищим, на рівні 27,9%, що вказує на високу придатність для систем з обмеженими обчислювальними ресурсами. Ці результати підтверджують практичність використання попередньо навчених моделей сегментації, після їх належної адаптації, для підтримки вимог розпізнавання перешкод у навігації в приміщенні для робототехнічних платформ.

Ключові слова: *обробка зображнь, розпізнавання шаблонів,сегментація зображень, тренування моделей.*

BORKIVSKYI BOHDAN, TESLYUK VASYL
Lviv Polytechnic National University

IMPROVING OBSTACLE RECOGNITION IN INDOOR ENVIRONMENTS FOR ROBOTIC SYSTEMS

Accurate obstacle recognition plays a crucial role in enabling robotic systems to operate safely and effectively in indoor environments, especially in automotive-scale applications where safety and reliability is essential. This study focuses on enhancing segmentation performance for common indoor obstacles, including furniture and structural elements. The study presents a focused review and analysis of recent advancements in segmentation foundation models, with a focus on their potential in robotic systems and the use of architectures such as SAM2. To assess their applicability, we adapt a general-purpose segmentation model by fine-tuning it on RGB images from the NYU Depth V2 dataset, which contains a wide range of indoor scenes captured from viewpoints similar to those encountered by robotic systems, depth information was excluded to reflect the limitations of monocular camera setups typical in embedded robotics. The adapted model achieves improved segmentation accuracy, reaching 82.4% mIoU, and demonstrates greater robustness in recognizing obstacles within cluttered and visually complex scenes. Experimental results show consistent improvements over the baseline, with a 7.6% gain in performance, supporting the value of domain-specific adaptation for improving robotic perception. Although segmentation quality varies with the number of point-based prompts (performance is worse with 5 points than with 15) the improvement over the baseline in the low-input case is considerably higher, at a level of 27.9%, indicating strong suitability for systems where computational resources are limited. These results confirm the practicality of using pre-trained segmentation models, once properly adapted, to support the demands of obstacle recognition in indoor navigation for automotive-scale robotic platforms.

Keywords: *image processing, pattern recognition, image segmentation, fine-tuning.*

Стаття надійшла до редакції / Received 07.12.2025
Прийнята до друку / Accepted 11.01.2026
Опубліковано / Published 29.01.2026

Introduction

Obstacle recognition is a fundamental component of robotic perception, directly influencing the safety, efficiency, and autonomy of mobile systems [1]. In indoor environments, this task becomes particularly complex due to cluttered scenes, varied object geometries, and frequent occlusions. For automotive-scale robotic systems designed to operate in constrained indoor settings-such as parking structures, warehouses, or industrial facilities-accurate obstacle segmentation is essential for reliable navigation and interaction.

While recent advances in visual segmentation, particularly those driven by foundation models trained on large-scale image corpora, have demonstrated strong generalization capabilities across diverse domains, their direct application to indoor obstacle recognition remains limited. Models trained primarily on outdoor or generic visual data often struggle to detect indoor obstacles such as furniture, fixtures, or structural elements with sufficient precision. Furthermore, existing datasets either over-rely on depth information or fail to reflect the complexity of real-world indoor environments encountered by autonomous robotic systems.

Despite their potential, the integration of segmentation-based foundation models into robotic systems remains at an early stage. Although preliminary research indicates that it is possible to adjust these models for specific tasks with competitive results [2], their effectiveness in robotic perception tasks - especially those requiring real-time processing, embedded systems, and varying sensors - needs additional exploration. Obstacle identification, a fundamental skill for safe and effective autonomous navigation, is one such activity. Identifying barriers requires not only great segmentation ability but also the capacity to recognize relevant aspects in diverse and chaotic contexts - often with incomplete or vague visual information [3].

The goal of this paper is to explore the practical use of recent advancements in segmentation foundation models in robotic systems. It offers a detailed summary of existing methods, identifies transferable techniques, and proposes a path forward for adapting these models to obstacle recognition. Through a thorough study of modern architectural trends, functional enhancements, and custom alterations, we aim to guide the following projects by integrating basic models as crucial components of robotic vision systems. In this process, we highlight both the potentials and the obstacles associated with synchronizing advanced visual inference systems with the needs of autonomous robotics in real-world applications.

Object of research – the process of object detection in a closed environment for robotic systems.

Subject of research – methods and means of object segmentation in an enclosed environment for robotic systems.

The purpose of the research – to increase the efficiency of object recognition in an enclosed environment for autonomous mobile systems with obstacle avoidance by finding clear boundaries of obstacles.

To achieve this purpose, the following main research objectives are identified:

Conduct a literature analysis of methods and means of object segmentation.

Choose the optimal architecture of a neural network for object segmentation.

Select a data set for training the neural network and develop a model.

Conduct training experiments with different combinations of model parameters.

Evaluate the accuracy of the selected algorithms.

Related Works

Obstacle recognition in indoor robotic systems has undergone a transformative evolution with the introduction of foundation vision models like Segment Anything Model 2 (SAM 2), which incorporates a hierarchical Vision Transformer encoder, multi-scale feature fusion, and promptable segmentation with support for point, box, mask, and textual prompts, alongside a lightweight temporal memory unit designed to enhance video segmentation consistency at rates approaching 44 fps on high-end GPUs, thus enabling real-time, class-agnostic object delineation in cluttered indoor environments [4]. SAM 2's architecture builds upon the groundwork of the original SAM and augments its capacity for promptable, zero-shot segmentation by operating effectively on both still images and video frames.

In robotics research, this capability has been extended through techniques like MoSAM, which enhances spatiotemporal coherence by injecting motion-guided prompts derived from optical flow and by implementing selective memory mechanisms to retain only temporally consistent frames, thus enabling superior tracking of occluded or fast-moving obstacles [5].

Simultaneously, multi-object tracking frameworks such as SAM2MOT leverage a tracking-by-segmentation paradigm, where temporal continuity is maintained via trajectory managers and cross-object interaction modules, achieving state-of-the-art performance across benchmarks like DanceTrack and UAVDT, and demonstrating that mask-based association can outperform detection-based paradigms by +2.1 HOTA and +4.5 IDF1 scores [6].

Parallel advances in related fields underscore SAM 2's adaptability: in surgical robotics, empirical evaluation of SAM 2 in video segmentation demonstrated robustness across various intraoperative scenarios; pathology imaging has employed Path-SAM 2 for adenoma detection in histopathological slides; and applications in knee MRI and retinal OCT segmentation have produced Dice scores north of 0.90 using zero-shot prompt strategies [7–9]. Beyond medical imaging, SAM 2 has been applied in articulated object manipulation, where it segments moving components in 3D point clouds - a critical enabler for axis estimation and precise grasping control in closed-loop robotic pipelines [10]. In environmental or agricultural domains, SAM 2-based frameworks have supported hyperspectral land classification and object detection in satellite or UAV imagery, demonstrating generalization across varied modalities [11].

Further, prototypes for embedding SAM 2 in interactive AR/VR pipelines and for Sonar image segmentation have illustrated the model's cross-domain versatility though highlighting persistent latency bottlenecks on embedded platforms [12]. Real-world deployment tests - such as robotic arms aiming to grasp dynamically moving balls - have revealed that full SAM 2 pipelines operate at only 3–7 fps on hardware like NVIDIA T4 or Jetson Orin, necessitating the development of lightweight variants (FastSAM, MobileSAM, EdgeSAM) and optimized I/O and prompt sampling strategies to uphold ≥20 fps throughput while preserving segmentation fidelity [13-15].

Complementary studies in embodied and embodied navigation have begun integrating SAM 2 into SLAM and costmap planners. For example, EmbodiedSAM, which enables online general-purpose 3D segmentation in indoor scenes, has been suggested as a means for enriching semantic mapping and dynamically updating cost layers for navigation [16]. Surveys of open-vocabulary semantic segmentation in embodied agents endorse the use of promptable foundation models to interpret arbitrary scene elements during navigation and to facilitate pedestrian and object-aware path planning [17].

Nevertheless, SAM 2 exhibits known limitations in real-time robotic settings - it struggles with tree-like or low-contrast textures, over-segmenting foliage or failing to separate subtly textured regions [18], and promptless segmentation may produce coarse boundaries that require refinement. Moreover, its original model lacks semantic awareness and must be paired with detectors (e.g., Grounding DINO, RT-DETR) to ground masks to known categories and facilitate detection-to-segmentation pipelines; community implementations like Grounded-SAM 2 demonstrate strong open-set semantic tracking capabilities via multi-model prompting [19]. In aggregate, this body of work highlights a clear technical trajectory: harness the generality and mask accuracy of SAM 2, enhance temporal consistency and motion-awareness through tailored prompt modules, embed lightweight variants into edge platforms, and integrate segmentation outputs into SLAM, costmap computation, and dynamic obstacle avoidance loops. The principal research gaps lie in reconciling high-accuracy, mask-level obstacle detection with real-time (<50 ms latency) performance on constrained robotic hardware; designing adaptive prompt strategies that leverage sensor fusion (e.g., odometry, depth) to optimize mask generation; and unifying obstacle segmentation with semantic labeling and tracking to enable context-aware navigation.

Materials and methods

In this study we focus on adapting a general-purpose segmentation foundation model to improve obstacle recognition in indoor environments, specifically targeting its application in robotic systems operating under RGB-only constraints. We selected Meta's Segment Anything Model 2 (SAM2) as the base model due to its balance of promptable segmentation capability and real-time performance, making it suitable for robotic applications where inference speed is critical.

Our goal was to enhance SAM2's segmentation accuracy for indoor obstacle classes such as tables, chairs, sofas, and cabinets - objects commonly encountered in automotive-scale robotics navigating confined indoor spaces. To achieve this, we fine-tuned the model using class-specific supervision and prompt signals in the form of sparse point annotations. The evaluation focused on assessing whether such targeted adaptation could significantly improve mask prediction quality in cluttered indoor scenes.

We considered two widely-used semantic segmentation datasets - ADE20K and NYU Depth V2. While ADE20K offers large-scale scene diversity and fine-grained semantic labels [20], it spans both indoor and outdoor domains and includes over 20,000 images. The complexity and heterogeneity of ADE20K, as well as the computational burden of sampling representative subsets, led us to favor the NYU Depth V2 dataset [3], which comprises 1,449 densely labeled indoor RGB-D images collected from 464 indoor scenes across a variety of settings such as offices, kitchens, and classrooms. Importantly, we discarded the depth modality and used only the RGB images and their corresponding semantic segmentation labels.

Fine-tuning was conducted in a prompted segmentation setting, consistent with SAM2's original interface. During each training step, we supplied a single random point located within a ground truth mask of an obstacle object class and the corresponding segmentation mask as the supervision target. This prompt-to-mask training strategy aligns with the core philosophy of foundation segmentation models - segment anything given a point or bounding box, while ensuring task specificity by enforcing supervision only on a fixed set of indoor obstacle categories. This strategy helped maintain consistency with the model's promptable inference design while aligning it with the obstacle recognition task.

The model was fine-tuned over 20,000 iterations using an AdamW optimizer, with a base learning rate of $1\times10{-}5$, a cosine annealing scheduler, and mixed-precision training for memory efficiency.

We adopted a controlled point-prompt evaluation setup to measure performance. At test time, we supplied the model with a single random point per ground truth object class, mirroring the conditions used during training. The predicted segmentation masks were then compared to the ground truth using the mean Intersection over Union (mIoU) metric, computed over the selected obstacle-relevant classes.

To ensure robustness, the evaluation was repeated over multiple random seeds for prompt selection, and the final reported mIoU reflects the average performance across these runs.

The fine-tuning of the segmentation model was conducted using the Google Colab environment, leveraging an NVIDIA Tesla T4 GPU with 16 GB of memory. Training was executed over a span of 21,000 iterations.

The training pipeline was implemented in PyTorch with mixed-precision enabled to optimize GPU utilization and memory efficiency. At each training iteration, a compound loss function was computed to guide the optimization process. This compound loss was composed of two key components - Binary Cross-Entropy (BCE) Loss and Mean Intersection over Union loss. BCE is applied between the predicted segmentation mask and the ground truth binary mask for the target object, this loss component ensured pixel-wise accuracy and penalized false positive and false negative classifications. mIoU is computed across all selected obstacle classes present in the batch. This term served to encourage overall segmentation consistency and improve performance on class-specific regions of interest. Unlike standard mIoU which is typically used only at evaluation, incorporating this metric during training helped the model optimize directly for the target evaluation criterion.

The total compound loss is formulated as a weighted sum:

$$\mathcal{L}_{total} = \mathcal{L}_{BCE} + 0.05(1 - mIoU), \quad (1)$$

where $\mathcal{L}_{BCE}$ is a binary cross entropy loss:

$$\mathcal{L}_{BCE} = -[y * \log(\hat{y}) + (1 - y) * log(1 - \hat{y})], \quad (2)$$

where y is the actual binary label, and $\hat{y}$ is the predicted probability of the positive class.

The inclusion of the (1-mIoU) term encouraged the model to improve global segmentation consistency, as it equally considers both precision and recall across object regions, and it penalizes both under- and over-segmentation while preserving pixel-level sensitivity through the BCE loss.

During training, the model's performance was monitored using a validation subset of the NYU Depth V2 dataset. The mIoU was computed at fixed intervals using point-prompt evaluation, where a single randomly selected point was provided per object class, mimicking the inference-time interaction paradigm. This approach ensured that evaluation remained consistent with the deployment scenario in robotics applications, where minimal user or system interaction is expected.

Results and discussion

The pre-fine-tuning baseline, representing the unadapted SAM2 model applied in a zero-shot fashion, achieved a mean IoU of 74.8% on the test set. This level of performance, while relatively strong, suffered from notable failure cases including fragmented masks for occluded objects, confusion between adjacent furniture classes, and over-segmentation in low-contrast regions.

Following fine-tuning, the model exhibited a consistent and measurable improvement across all evaluated classes. After 21,000 training iterations, the mean IoU increased to 82.4%, representing a 7.6 percentage point absolute gain over the baseline. This improvement validates the effectiveness of domain-specific adaptation. Table 1 summarizes the improvements in recognition quality at different stages of fine-tuning:

Table 1

mIoU score based on number of fine-tune iterations

Iterations passed	mIoU (%) based on number of points per class		
	5	10	15
0 (not fine-tuned model)	41.9	67.6	74.8
7000	60.4	71.9	77.2
14000	66.1	78.8	80.5
21000	69.8	80.5	**82.4**

To further confirm that fine-tuning substantially improved spatial coherence and class-consistent segmentation we performed visual inspection of sample outputs.

In qualitative comparisons, the fine-tuned model showed enhanced robustness to background clutter and complex lighting conditions. For instance, in office-like environments (Fig. 1, Fig. 2) with partially occluded chairs and overlapping furniture, the adapted model consistently produced more complete and contextually accurate segmentation masks. Improvements were also visible in boundary alignment, particularly for irregular object geometries. Although there were difficulties with segmenting background objects, qualitative inspection of segmentation outputs revealed that foreground objects, particularly those in close proximity to the camera, were segmented with noticeably higher precision compared to background objects. This is a favorable characteristic for obstacle recognition systems, where accurate delineation of nearby objects is crucial for real-time navigation and collision avoidance. Additionally, the floor regions were consistently segmented with high accuracy using fine-tuned model. This reliability in floor segmentation contributes to a more stable estimation of the room's geometric layout, aiding downstream modules such as path planning and free-space mapping in robotic systems.

Fig. 1. Recognizing obstacles in the office:
a) baseline result; b) input image; c) fine-tuned result

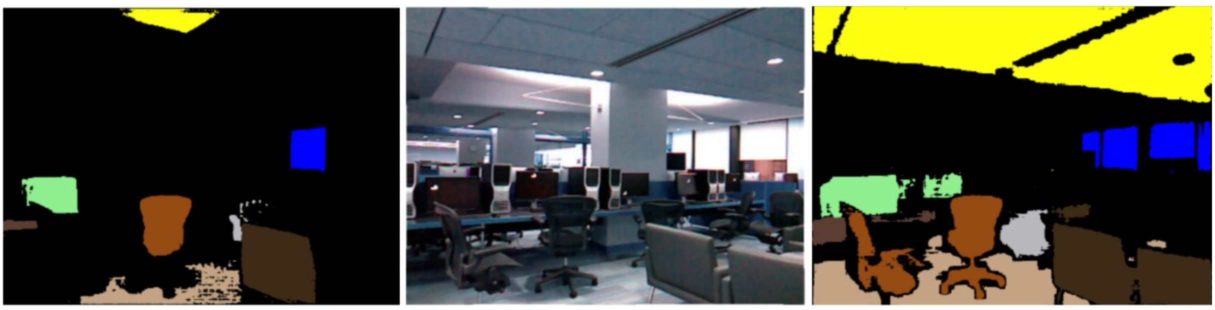

Fig. 2. Recognizing obstacles in the classroom:
a) baseline result, b) input image, c) fine-tuned result

The combination of improved foreground segmentation and consistent floor detection highlights the model's capacity to capture both object-level and spatial-context cues, which are critical for safe and efficient robotic operation in cluttered indoor environments (Fig. 3). Baseline model was able to correctly segment only part of floor and failed to locate other part of aisle due to low lighting conditions, in contrast fine-tuned model was able to correctly segment all furniture as well as floor, providing full understanding of the surrounding in the given scene, further confirming that adapting a general-purpose segmentation model for indoor environments operated by robotic systems can result in much better segmentation quality.

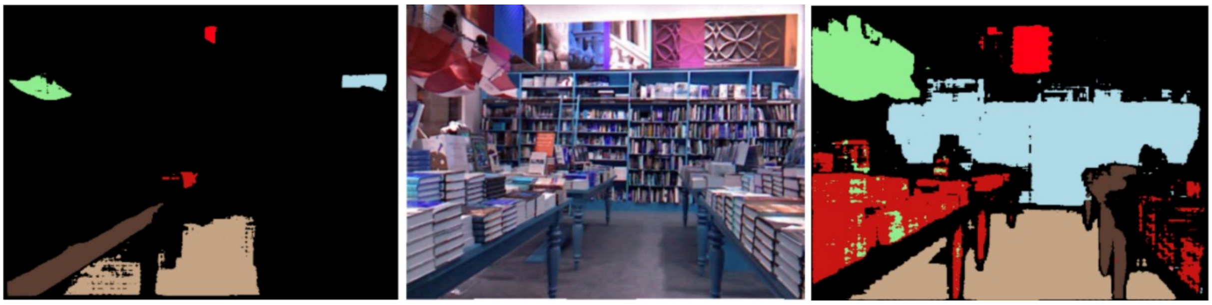

Fig. 3. Recognizing obstacles in the cluttered scene:
a) baseline result, b) input image, c) fine-tuned result

Thus, based on the results of the work performed, we can formulate the following scientific novelty and practical significance of the research results.

Scientific novelty of the obtained research results – that the method of object recognition using neural networks has been improved by fine-tuning foundation segmentation model with relevant data leading to improvements in obtaining accurate boundaries of objects in the image, that are supported by analysis of quantitative and qualitative results.

Practical significance of the research results – that the improved method of object recognition can be used in robotic systems for social purposes, for example, in robot assistants or automotive delivery robots, or in military domain, for example, during demining in various conditions to detect dangerous objects or during rescue and search operations.

Conclusions

In this work, we investigated the adaptation of a foundation segmentation model for the task of obstacle recognition in indoor environments relevant to robotic systems. By fine-tuning SAM2 on the NYU Depth V2 dataset using only RGB data and prompt-based supervision with sparse point annotations we demonstrated that general-purpose segmentation models can be effectively specialized for robotics applications with minimal modification.

Our training approach, which combined binary cross-entropy loss with a mean IoU-based objective, resulted in a significant performance gain - the model's mean Intersection over Union improved from 74.8% to 82.4% after 21,000 training iterations. This improvement reflects enhanced segmentation accuracy for key obstacle classes such as chairs, tables, and cabinets, particularly in cluttered or occluded indoor scenes. The model also maintained consistent and accurate segmentation of floor surfaces, which is essential for spatial reasoning in navigation tasks.

Qualitative analysis further confirmed that close-range objects were segmented with high precision - an essential capability for real-time obstacle detection. The consistent delineation of the floor across scenes also supports downstream robotic functions such as environment mapping and safe trajectory planning.

However, some limitations were observed. Segmentation quality declined for smaller or partially occluded objects in the background, and fine-grained class differentiation was occasionally imprecise in densely populated scenes. These weaknesses suggest that while fine-tuning greatly enhances near-field perception, additional strategies, such as multi-scale feature refinement or semantic priors, may be needed to improve performance in more visually complex or distant areas of the scene.

References

1. Tsmots, I., et al, "Development of the scheme and improvement of the motion control method of a group of mobile robotic platforms", Ukrainian Journal of Information Technology, 5(2), 97 104, 2023. https://doi.org/10.23939/ujit2023.02.097.

2. Kuan-Lin Chen, et al, "Resolving Positional Ambiguity in Dialogues by Vision-Language Models for Robot Navigation," 2024. https://doi.org/10.48550/arXiv.2410.12802.

3. Silberman, N., et al, "Indoor Segmentation and Support Inference from RGBD Images," in Computer Vision – ECCV 2012, 2012, pp. 746–760. https://doi.org/10.1007/978-3-642-33715-4_54.

4. Nikhila Ravi, et al, "SAM 2: Segment Anything in Images and Videos," 2024. https://doi.org/10.48550/arXiv.2408.00714.

5. Qiushi Yang, et al, "MoSAM: Motion-Guided Segment Anything Model with Spatial-Temporal Memory Selection," 2025. https://doi.org/10.48550/arXiv.2505.00739.

6. Junjie Jiang, et al, "SAM2MOT: A Novel Paradigm of Multi-Object Tracking by Segmentation," 2025.

https://doi.org/10.48550/arXiv.2504.04519.
7. Jieming Yu, et al, "SAM 2 in Robotic Surgery: An Empirical Evaluation for Robustness and Generalization in Surgical Video Segmentation," 2024. https://doi.org/10.48550/arXiv.2408.04593.
8. Mingya Zhang, et al, "Path-SAM2: Transfer SAM2 for digital pathology semantic segmentation," 2024. https://doi.org/10.48550/arXiv.2408.03651.
9. Xinrun Chen, et al, "SAM-OCTA: Prompting Segment-Anything for OCTA Image Segmentation," 2024. https://doi.org/10.48550/arXiv.2310.07183.
10. Xi Wang, et al, "Articulated Object Manipulation using Online Axis Estimation with SAM2-Based Tracking," 2025. https://doi.org/10.48550/arXiv.2409.16287.
11. Xue, B., et al. "Adapting Segment Anything Model to Aerial Land Cover Classification With Low-Rank Adaptation," in IEEE Geoscience and Remote Sensing Letters, vol. 21, pp. 1-5, 2024. http://doi.org/10.1109/LGRS.2024.3357777.
12. Lin Wang, et al, "When SAM Meets Sonar Images," 2023. https://doi.org/10.48550/arXiv.2306.14109.
13. Xu Zhao, et al, "Fast Segment Anything," 2023. https://doi.org/10.48550/arXiv.2306.12156.
14. Chaoning Zhang, et al, "MobileSAMv2: Faster Segment Anything to Everything," 2023. https://doi.org/10.48550/arXiv.2312.09579.
15. Chong Zhou, et al, "EdgeSAM: Prompt-In-the-Loop Distillation for On-Device Deployment of SAM," 2024. https://doi.org/10.48550/arXiv.2312.06660.
16. Xiuwei Xu, et al, "EmbodiedSAM: Online Segment Any 3D Thing in Real Time," 2025. https://doi.org/10.48550/arXiv.2408.11811.
17. Chunhui Zhang, et al, "A Comprehensive Survey on Segment Anything Model for Vision and Beyond," 2023. https://doi.org/10.48550/arXiv.2305.08196.
18. Yixin Zhang, et al, "Quantifying the Limits of Segmentation Foundation Models: Modeling Challenges in Segmenting Tree-Like and Low-Contrast Objects," 2025.https://doi.org/10.48550/arXiv.2412.04243.
19. Tianhe Ren, et al, "Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks," 2024. https://doi.org/10.48550/arXiv.2401.14159.
20. Zhou, B., et al, "Scene Parsing through ADE20K Dataset," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 5122-5130. https://doi.org/10.1109/CVPR.2017.544.