Науковий вісник НЛТУ України
Scientific Bulletin of UNFU
ISSN 1994-7836 (print)
ISSN 2519-2477 (online)
https://doi.org/10.36930/40360117
Article received: 15.12.2025, accepted: 06.02.2026
Article published: 26.02.2026
UDC 004.[93+4]
Corresponding author: B. P. Borkivskyi bohdan.p.borkivskyi@lpnu.ua

B. P. Borkivskyi, V. M. Teslyuk
Lviv Polytechnic National University, Lviv, Ukraine

SPECIALIZED STRUCTURAL PRUNING OF FOUNDATION MODELS FOR INDOOR ROBOTIC VISION

We address the problem of deploying large-scale vision foundation models in real-time robotic systems, which is currently hindered by immense computational demands and significant inference latency. We analyze the limitations of general-purpose segmentation model compression methods, which frequently fail to maintain semantic precision and mask granularity when transferred to specialized environments such as indoor robotic navigation. We introduce a domain-specific structural optimization framework for segmentation models, designed to transform a heavyweight image segmentation architecture into a lightweight, high-precision tool tailored for specialized perception tasks. We examine the role of a multi-stage pipeline, beginning with a critical adaptation phase where a full-scale teacher model is specialized to establish robust prior knowledge of target textures and indoor geometries, providing a domain-aware baseline for knowledge transfer. We implement an alternate slimming framework that decomposes the vision transformer encoder into decoupled embedding and bottleneck internal dimensions, allowing sequential structural pruning that maintains architectural consistency for intermediate feature alignment. We integrate a robust data augmentation pipeline during the distillation and recovery phases, incorporating complex geometric and photometric transformations to stabilize the learning process and mitigate overfitting in data-scarce scenarios. Our results show that this comprehensive approach achieves a 73.3 % reduction in total trainable parameters and a 74.3 % reduction in Multiply-Accumulate operations (MACs), which effectively doubles the inference speed from 7 FPS to 15 FPS. We observe that the optimized student model experiences a marginal accuracy drop of only approximately 1 % compared to the specialized teacher model, consistently outperforming standard SlimSAM baselines in both semantic mask granularity and overall environmental robustness. We establish that our model demonstrates superior resistance to lighting variations and shadow interference, successfully addressing a common failure point in baseline architectures like FastSAM, that often distort segmentation masks under high-contrast indoor illumination. We conclude that the integration of domain-specific pruning and augmented distillation provides a superior and more stable path for enabling real-time, sophisticated computer vision on resource-constrained autonomous platforms.

Keywords: image processing; data processing; image segmentation; fine-tuning; general-purpose segmentation models.

Introduction / Вступ

The ability to accurately perceive and interpret visual information is a fundamental requirement for modern autonomous systems [14]. Object detection and segmentation serve as the cornerstone of this perception, enabling machines to not only identify the presence of various entities but also to delineate their precise boundaries. This level of granularity is especially critical in the domain of robotic systems, where robots must interact with complex, unstructured environments in real-time. Whether it is a robotic arm performing precise manipulation or an autonomous mobile unit navigating a crowded indoor space, the system's safety and efficiency depend on its ability to distinguish between objects, obstacles, and surfaces with high reliability.

The emergence of foundation models, such as the Segment Anything Model (SAM), has revolutionized this field by providing a unified framework for class-agnostic segmentation. These models are trained on massive datasets comprising millions of images, allowing them to demonstrate extraordinary zero-shot generalization capabilities [3]. For robotics, this means a pre-trained model can often segment unfamiliar objects without the need for extensive task-specific retraining, significantly lowering the barrier for deploying versatile autonomous agents in diverse scenarios.

However, despite their impressive accuracy, these foundation models face significant limitations when applied to practical, real-world tasks [20]. Their most prominent drawback is their immense computational complexity and massive parameter counts. The heavyweight architectures required to achieve high generalization result in significant inference latency and high memory consumption. In the context of robotics, where decisions must often be made in milliseconds and hardware resources are limited by power and size constraints, the deployment of such "dense" networks remains largely impractical.

To overcome these limitations, this research focuses on the optimization of the network's internal structure through a process of strategic reduction. Rather than replacing the

Борківський Богдан Петрович, магістр, аспірант, кафедра автоматизовані системи управління.
Email: bohdan.p.borkivskyi@lpnu.ua; https://orcid.org/0000-0003-3301-476X
Теслюк Василь Миколайович, д-р техн. наук, професор, завідувач кафедри автоматизовані системи управління.
Email: vasyl.m.teslyuk@lpnu.ua; https://orcid.org/0000-0002-5974-9310
Цитування за ДСТУ: Борківський Б. П., Теслюк В. М. Спеціалізоване структурне спрощення базових моделей для систем технічного зору роботів у приміщеннях. Науковий вісник НЛТУ України. 2026, т. 36, № 1. С. 156–161.
Citation APA: Borkivskyi, B. P., & Teslyuk, V. M. (2026). Specialized structural pruning of foundation models for indoor robotic vision. *Scientific Bulletin of UNFU*, 36(1), 156–161. https://doi.org/10.36930/40360117

architecture entirely, we can streamline the existing model by identifying and removing redundant components while retaining its core knowledge. By employing a progressive compression framework, the model is thinned in stages, allowing it to maintain its sophisticated understanding of visual features while becoming significantly faster and more lightweight. This methodology allows the adaptation of powerful foundation models into efficient, domain-specific tools, making high-precision object recognition viable for resource-constrained robotic applications.

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

Subject of research – methods and means of object segmentation for robotic systems in an enclosed environment, that allow improving the performance of segmentation model and speed of inference.

The purpose of the research – to increase the performance of object segmentation model for robotic systems in an enclosed environment by applying structural pruning, that will allow using such model in resource-constrained systems.

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

- conduct a literature analysis of methods and means of object segmentation, that will allow systematization of existing approaches to accelerating vision foundation models and the identification of critical gaps in adapting these architectures for robotic systems within specialized domains;
- choose the optimal method of architecture slimming for object segmentation neural networks, that will enable deep compression of vision transformer backbones without compromising structural integrity;
- select a data set for training the neural network and develop a model, that will facilitate the implementation of a domain adaptation phase for transferring knowledge of indoor-specific geometries and textures;
- conduct training experiments with different combinations of model parameters, that will allow the exploration of the trade-offs between computational complexity and inference throughput;
- evaluate the accuracy of the selected algorithms, that will verify the effectiveness of the proposed methodology against existing solutions.

Analysis of recent research and publications. The advent of Vision Foundation Models (VFMs) has revolutionized image segmentation by moving away from task-specific training toward broad-domain generalization. The Segment Anything Model (SAM) [12] is the leading framework in this space, having been trained on the massive SA-1B dataset [16], which consists of 11 million high-resolution images and over 1 billion segmentation masks.

SAM's architecture is split into three main parts: a Vision Transformer (ViT) image encoder [5], a prompt encoder, and a mask decoder. While the prompt encoder and mask decoder are highly efficient, the ViT image encoder is the primary source of computational overhead due to the quadratic complexity of its self-attention mechanism. This heavyweight encoder contributes to over 95 % of the model's parameter count and total MACs (Multiply-Accumulate Operations), creating a significant bottleneck for real-time deployment.

Model pruning [6] is a fundamental technique for eliminating parameter redundancy within deep neural networks. Research generally distinguishes between unstructured pruning, which removes individual weights and requires specialized hardware, and structural pruning, which eliminates entire architectural components such as channels, filters, or layers. Structural pruning is particularly effective for achieving real-time speedups on standard CPUs and GPUs without specialized software.

The success of structural pruning depends on importance estimation (saliency) [17], the process of identifying which parameters contribute least to the model's output. Common criteria include magnitude-based pruning or first-order Taylor expansions, which approximate the loss change caused by removing a specific parameter. In complex architectures like Transformers, pruning must also account for structural dependencies, where changing the dimension of one layer necessitates changes in others, a challenge addressed by dependency tracking frameworks like DepGraph.

Knowledge Distillation (KD) [7] serves as a vital tool for recovering the performance of a model after it has been thinned by pruning. The process involves a "teacher" model (the original dense SAM) supervising a "student" model (the pruned lightweight version). Beyond using final soft targets to transfer knowledge, state-of-the-art distillation often incorporates intermediate feature alignment.

FastSAM [21] approach reformulates the "segment anything" task as an all-instance segmentation problem followed by prompt-guided selection. It utilizes a YOLOv8-seg detector, which is a CNN-based architecture optimized for real-time performance. By leveraging the local inductive biases of convolutions, FastSAM achieves up to a 50 x speedup compared to the original SAM-H during inference. However, as noted in the research, this speed comes with a trade-off in mask granularity, particularly for small objects.

Recognizing the bottleneck of the image encoder, MobileSAM [19] adopts a "decoupled" distillation approach. It replaces the heavy ViT encoder with a TinyViT backbone. By distilling knowledge from the original SAM encoder into this lightweight architecture, MobileSAM achieves significantly reduced parameter counts and computational demands suitable for mobile applications.

Diverging from methods that require training new architectures from scratch, SlimSAM [4] focuses on reusing the robust prior knowledge of pre-trained SAMs through structural optimization, which is achieved through several pioneering contributions. Instead of a single-step aggressive pruning, alternate slimming framework decomposes the model into two separate sub-structures: the embedding (output dimensions of each block) and the bottleneck (intermediate features of each block). By sequentially pruning and restoring these components in an alternating fashion, SlimSAM prevents the steep performance degradation typically associated with high compression ratios. To enhance performance recovery, the intermediate feature alignment exploits the hidden state information of the original model. By partitioning the model into sub-structures, SlimSAM circumvents dimensionality inconsistency issues, enabling the pruned model to align its intermediate feature maps with the original model through consistent backpropagation.

The refinement of how parameter importance [10] is calculated is a critical innovation in structural reduction. Traditional Taylor-order approximations often suffer from a misalignment between the pruning objective (minimizing hard label discrepancy) and the distillation target (minimizing soft label loss). The disturbed Taylor importance method calculates gradients based on the loss between the ori-

ginal image embedding and a "disturbed" embedding injected with Gaussian noise, resulting in label-free estimation.

The reviewed literature illustrates a rapid evolution in foundational model acceleration, ranging from high-speed architectural substitutes to data-efficient slimming frameworks. However, a critical deficiency remains in the systematic application of these methods to domain-specific robotic perception: current approaches do not sufficiently address the necessity of preliminary domain adaptation before structural reduction, the impact of environmental volatility, such as shadow interference and dynamic lighting, on pruned weights is not thoroughly quantified, and the efficacy of augmented knowledge inheritance in data-scarce scenarios has not been formalized. This gap motivates the present study, which focuses on a specialized multi-stage pipeline for alternate slimming. By identifying the optimal trade-off between architectural sparsity and domain-specific accuracy, this research clarifies the performance-efficiency requirements necessary to transform heavyweight foundation models into viable perception tools for computationally constrained robotic systems.

Materials and methods of research. In this research we investigate optimizing segmentation foundation models under resource-constrained conditions. The study is conducted with the usage of NYU Depth v2 imagery dataset [13] that consist of diverse indoor scenes as the primary source for domain adaptation and slimming with no additional depth information utilized, ensuring the model relies solely on visual semantic features for object recognition. Primary metric used for accuracy – mAP, for efficiency – parameters count, MAC and fps. All training was performed in Google Colaboratory environment with NVIDIA A100 GPU.

Research results and their discussion / Результати дослідження та їх обговорення

Domain-specific pre-training of image segmentation model. The effectiveness of structural pruning is heavily dependent on the quality of the "teacher" model from which the pruned "student" inherits its knowledge. While foundation models like the Segment Anything Model (SAM) are trained on massive datasets like SA-1B to achieve open-world generalization, they often lack the specialized feature representations required for high-precision tasks in niche environments, such as indoor robotic navigation.

Before initiating the structural reduction process, it is essential to establish a robust domain-aware baseline. We implemented a preliminary fine-tuning phase, where the original, dense SAM-B (Base) model was trained on the NYU Depth RGB data [1]. This step shifts the model's focus from general-purpose segmentation to the specific textures, lighting conditions, and object categories common in indoor environments. By fine-tuning the full-scale model, we create a specialized teacher that possesses "domain-specific robust prior knowledge". This specialized teacher provides more relevant intermediate feature maps for the student model to align with during the subsequent distillation stages, particularly when training data is severely limited.

Alternate slimming framework. To achieve high compression ratios while maintaining the model's representational integrity, the alternate slimming framework [4] decomposes the complex internal structure of the Segment Anything Model (SAM) into decoupled components for sequential optimization. This methodology prevents the steep performance degradation typically associated with aggressive, single-step pruning by minimizing the architectural divergence between the original dense model and the pruned student.

An important component of the structural reduction process is the identification of redundant parameters using the disturbed Taylor importance criterion. Instead of relying on hard labels, gradients are calculated based on the loss between the original image embeddings and "disturbed" embeddings injected with Gaussian noise. This creates a label-free framework that seamlessly aligns the pruning criteria with the optimization targets of the distillation process, thereby enhancing accuracy recovery in data-limited scenarios.

The distillation process utilizes specific loss functions designed to maximize knowledge retention through intermediate feature alignment.

Bottleneck Aligning Loss (L_{Bn}). During the first distillation phase, the objective is to synchronize the dimensionality-consistent bottleneck features H and the final image embeddings *t*:

$$L_{Bn} = \alpha \cdot L_{MSE}\left(H_{v0}, H_{v1}\right) + \left(1-\alpha\right) \cdot L_{MSE}\left(t_{v0}, t_{v1}\right),$$

where H_{v0} is the teacher encoder intermediate features and H_{v1} is the encoder intermediate features after embedding pruning, t_{v0} is the final image embeddings from the original encoder, and t_{v1} is the final image embeddings from the first pruned encoder, L_{MSE} denotes the Mean-Squared Error, and α is a dynamic weight set to 0.5 for the first 10 epochs and 0 thereafter, allowing the model to focus purely on the final embedding after initial feature stabilization.

Embedding Aligning Loss (L_{Emb}). In the final phase, the loss function is structured to facilitate a smooth transition of knowledge from both the original teacher and the intermediate thinned model:

$$L_{Emb} = \alpha \cdot \left(L_{MSE}(E_{v1}, E_{v2}) + L_{MSE}(t_{v1}, t_{v2})\right) + \left(1-\alpha\right) \cdot L_{MSE}(t_{v0}, t_{v2}),$$

where E_{v1} is the first pruned encoder output features, E_{v2} is the output features of final pruned encoder, t_{v0} is the image embeddings from the original encoder, t_{v1} is the image embeddings from the first pruned encoder, and t_{v2} is the image embeddings from the final pruned encoder. The dynamic weight

$$\alpha = \frac{N-n-1}{N},$$

where $N = 10$, so that it progressively diminishes to zero as distillation unfolds. This transition shifts the learning objective from the thinned intermediate state $v1$ back to the original full-scale teacher $v0$, ensuring the highest possible precision in the final lightweight model.

Data augmentation. During the distillation and recovery phases, we observed that training on the limited NYU Depth subset led to poor results due to the model's inability to generalize from a static distribution. To overcome this, we implemented a robust data augmentation pipeline. It included geometric transformations, like random cropping, horizontal flipping, and scaling that were applied to increase spatial diversity and photometric adjustments, like color jittering and brightness variations that were used to simulate diverse indoor lighting conditions.

The inclusion of data augmentation proved essential [2] for stabilizing the distillation process, preventing the "steep performance degradation" often seen in data-scarce scenarios. It allowed the pruned student model to inherit the teacher's robust prior knowledge more effectively while adapting to the specific variance of indoor scenes.

Quantitative results of model pruning. By utilizing our proposed domain-specific optimization pipeline, which integrates an initial fine-tuning phase on the NYU Depth dataset with a data-augmented alternate slimming process, we achieve structural reduction with almost no performance degradation (Table 1). Specifically, our results show that the optimized model experiences a marginal accuracy drop of approximately 1 % compared to the full-scale teacher model. This minimal loss demonstrates that our methodology successfully inherits and preserves the core domain-specific knowledge established during pre-training, enabling the pruned student model to match the fidelity of a dense architecture while operating with far greater efficiency.

Table 1. Training results / Результати тренування моделі

Model	mAP (%)
Teacher	71.38
Sliming result	70.32

In terms of architectural scale, the structural reduction achieved through the alternate slimming framework was substantial (Table 2). The model was thinned from a baseline of 89,578,240 parameters to 23,906,248 parameters, representing a reduction of approximately 73.3 %. This drastic decrease in parameter count was matched by a corresponding decline in computational complexity, with total MACs dropping from 368.86 G to 94.93 G. This nearly 74.3 % reduction in Multiply-Accumulate Operations highlights the framework's ability to identify and eliminate massive redundancy within the Vision Transformer blocks, specifically within the decoupled bottleneck and embedding dimensions.

The impact on real-time performance is equally profound, with the model's inference speed effectively doubling from 7 FPS to 15 FPS. This 2.14x acceleration represents a critical threshold for robotic systems, moving the model from a latent, "stop-and-process" state toward the responsiveness required for continuous autonomous navigation. Unlike architectural replacement methods such as FastSAM, which achieve high speeds by substituting Transformers with CNNs, often at the cost of mask granularity, our method achieves competitive speeds while maintaining the sophisticated representational power of the original Transformer-based backbone.

Table 2. Efficiency metrics comparison with other models / Порівняння метрик ефективності з іншими моделями

Model	Dataset	Params (M)	MACs	Fps
SAM-B	SA-1B	90	369	7
SlimSAM	SA-1B	26	98	14
Ours	NYU	24	95	15

Qualitative comparison of image segmentation. A direct qualitative comparison between the domain-adapted teacher and the resulting optimized student models (Fig 1) further validates the efficacy of our structural reduction approach. Across the majority of indoor scenes in the NYU Depth dataset, the segmentation masks produced by the two models are virtually identical, which is consistent with the marginal 1 % accuracy drop recorded in our quantitative evaluation. This visual similarity indicates that the core spatial reasoning and feature extraction capabilities of the dense teacher model were successfully preserved throughout the alternate slimming process.

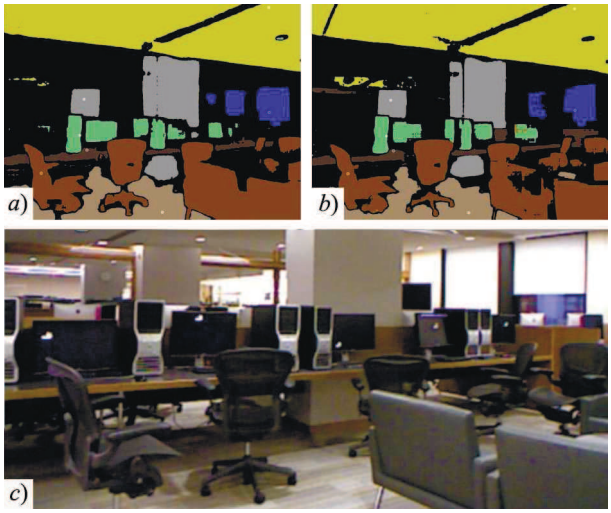

Fig. 1. Comparison of models image segmentation results / Порівняння результатів сегментації зображень моделей: *a*) baseline teacher model / початкова модель; *b*) input image / вхідне зображення; *c*) result model / отримана модель

Notably, for specific object categories such as computers and electronic peripherals, the optimized model demonstrates a surprising enhancement in mask fidelity. In several instances, the student model produces even more accurate and refined masks than its teacher. This phenomenon suggests that the structural reduction process, combined with targeted data augmentation, may serve as a form of architectural regularization. By pruning redundant parameters and focusing on the most salient features of the indoor domain, the optimized network effectively filters out "noise" from the teacher's broader representations, resulting in sharper edges and better-defined object boundaries for complex geometric shapes.

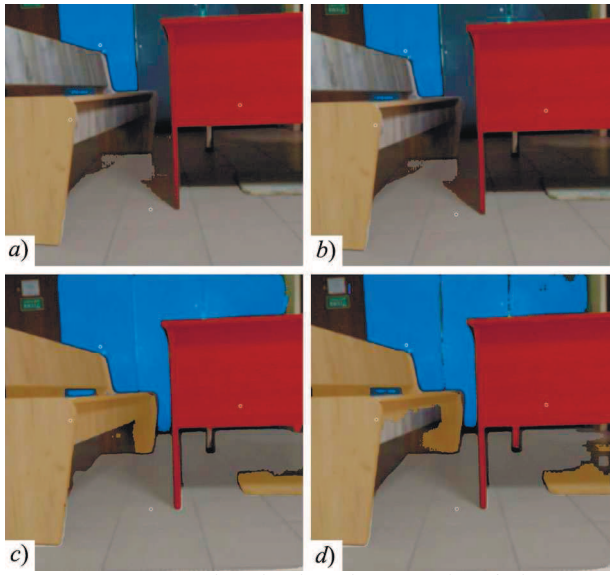

Fig. 2. Image segmentation visual results comparison / Візуальне порівняння результатів сегментації зображень: *a*) SAM-B / SAM-B; *b*) SlimSam / SlimSam; *c*) teacher model / початкова модель; *d*) result model / отримана модель

A visual analysis of the segmentation masks reveals distinct differences in how each model handles the complexities of indoor environments (Fig 2). While the original SAM-B baseline and the SlimSAM often produce results that are comparable to each other in general scenarios, they both struggle with the specific challenges of indoor lighting. In contrast, our proposed model demonstrates superior accu-

racy and significant robustness against adverse environmental factors.

Our optimized model, however, is not affected by lighting conditions in the same manner. By utilizing a domain-specific slimming pipeline, our model has learned to distinguish between semantic object boundaries and transient photometric changes. Consequently, our model produces more accurate and reliable results even in poorly lit areas or scenes with high-contrast shadows where other models fail, with better slimming rate and accuracy preservation.

Discussion of research results. The baseline SAM-B [8] exhibits sensitivity to illumination changes; specifically, its segmentation performance is frequently affected by shadows. In this case, the model may either fail to detect objects obscured by shadows or incorrectly merge shadow regions into the object masks, leading to distorted boundaries.

Model slimming can be achieved by applying token pruning [9] using an attention-based multilayer network. Although this approach results in faster ViT training process, when compared to other pruning techniques, in some cases accuracy drop is at a level of 2 %, while performance boost in terms of inference speed is only 35 %. Contrary, our result doubles inference speed with a smaller 1 % accuracy drop.

Although pruning CNNs for semantic segmentation on FPGA [11] solves the problem of accelerating a state-of-the-art semantic segmentation network by an FPGA-based accelerator exploiting a HW-aware channel pruning technique, but the resulting architecture is slimmed by 63 % and accuracy dropped by 1.5 %, while our method demonstrated 73 % slimming of segmentation model and just 1 % accuracy drop, thus achieving better slimming results.

SlimSAM [4] reaches similar efficiency level to the one of our segmentation model, with 71 % reduction in terms of model structure size and nearly the same speed of inference at 14 fps. Masks produced by this model are less spiky, when compared to baseline segmentation model, but it fails to work properly under hard lighting conditions as well.

Since Faster Sam uses different internal encoder architecture (YOLO instead of SAM) [18], its results are not affected by lighting conditions and shadows, but it exhibited poor boundary localization for thin structures, frequently missing short table legs entirely. It is also prone to confusing different objects of similar color, like confusing doors with furniture, or missing obstacles laying on the floor.

The discussed results prove the importance and necessity of our study, since none of the other methods can provide high accuracy for indoor object segmentation while applying model pruning. By utilizing a domain-specific teacher adapted to indoor textures and incorporating a robust data augmentation pipeline during distillation, the resulting model reaches high level of pruning without compromising on segmentation quality even in harsh lighting conditions.

So, based on the results of the work performed, it is possible to formulate the following scientific novelty and practical significance of the research results.

Scientific novelty of the obtained research results – improved the method of alternate slimming of segmentation foundation models by integrating data augmentation step, that leads to better preservation of segmentation accuracy on small-scale datasets.

Practical significance of the research results – the resulting low-parameter image segmentation model can be used in robotic systems with near-realtime requirements, like low-speed robots for social purposes, or in military domain during demining in various conditions to detect dangerous objects.

Conclusions / Висновок

Increased the performance of object segmentation model for robotic systems in an enclosed environment by applying structural pruning, that allows using such model in resource-constrained systems. Based on the results of the research the following main conclusions can be drawn.

1. Applied an alternate slimming framework to sequentially prune decoupled embedding and bottleneck dimensions, that allowed consistent dimensionality during distillation, facilitating superior knowledge transfer through intermediate hidden state alignment.
2. Integrated data augmentation step during alternate slimming step, that proved to be vital for stabilizing the recovery process, enabling the model to overcome environmental challenges such as shadow interference and varied lighting conditions that often affect baseline architectures.
3. Demonstrated the practical impact of this approach with a quantitative results analysis. The optimized model achieved a 73.3 % reduction in parameters and a 74.3 % reduction in MACs, effectively doubling the inference speed from 7 FPS to 15 FPS. Crucially, these efficiency gains were realized with only a marginal 1 % accuracy drop from the specialized teacher model, a performance level that matches the efficiency-accuracy standards of the original SlimSAM framework.
4. Our work proved that the internal structural optimization of transformer-based models, when guided by domain-specific priors and robust training strategies, is a superior path for enabling real-time, high-precision robotic perception.

References

1. Borkivskyi, B. P., & Teslyuk, V. M. (2026). Improving obstacle recognition in indoor environments for robotic systems. Herald of Khmelnytskyi National University. *Technical sciences*, 1-2026 (pp. 135–140). https://doi.org/10.31891/2307-5732-2025-359-17
2. Che, Q., Le, D., Pham, B., Lam, D., & Nguyen, V. (2025). Enhanced Generative Data Augmentation for Semantic Segmentation via Stronger Guidance. In *Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods ICPRAM*, vol. 1 (pp. 251–2). https://doi.org/10.5220/0013175900003905
3. Chen, K., Wei, T., Yeh, L., Kao, E., Tseng, Y., & Chen, J. (2024). Resolving Positional Ambiguity in Dialogues by Vision-Language Models for Robot Navigation. arXiv:2410.12802v1. https://doi.org/10.48550/arXiv.2410.12802
4. Chen, Z., Fang, G., Ma, X., & Wang, X. (2024). SlimSAM: 0.1 % Data Makes Segment Anything Slim. arXiv:2312.05284v4. https://doi.org/10.48550/arXiv.2312.05284
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. arXiv:2010.11929v2. https://doi.org/10.48550/arXiv.2010.11929
6. Fang, G., Ma, X., Song, M., Mi, M. B., & Wang, X. (2023). DepGraph: Towards Any Structural Pruning. arXiv:2301.12900v2. https://doi.org/10.48550/arXiv.2301.12900
7. Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. arXiv:1503.02531v1. https://doi.org/10.48550/arXiv.1503.02531
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W., Dollár, P., & Girshick, R. (2023). Segment Anything. arXiv:2304.02643v1. https://doi.org/10.48550/arXiv.2304.02643

9. Marchetti, M., Traini, D., Ursino, D., & Virgili, L. (2025). Efficient token pruning in Vision Transformers using an attention-based Multilayer Network. Expert Systems with Applications, 279, article ID 127449. https://doi.org/10.1016/j.eswa.2025.127449
10. Molchanov, P., Mallya, A., Tyree, S., Frosio, I., & Kautz, J. (2019). Importance Estimation for Neural Network Pruning. arXiv:1906.10771v1. https://doi.org/10.48550/arXiv.1906.10771
11. Morì, P., Vemparala, M.R., Fasfous, N., Mitra, S., Sarkar, S., Frickenstein, A., Frickenstein, L., Helms, D., Nagaraja, N., Stechele, W., & Passerone, C. (2022). Accelerating and pruning CNNs for semantic segmentation on FPGA. In *Proceedings of the 59th ACM/IEEE Design Automation Conference* (pp. 145–150). Association for Computing Machinery. https://doi.org/10.1145/3489517.3530424
12. Ravi, N., Gabeur, V., Hu, Y., 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., Girshick, R., Dollár, P., & Feichtenhofer, C. (2024). SAM 2: Segment Anything in Images and Videos. arXiv:2408.00714v2. https://doi.org/10.48550/arXiv.2408.00714
13. Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor Segmentation and Support Inference from RGBD Images. In *Computer Vision – ECCV 2012* (pp. 746–760). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-33715-4_54
14. Tsmots, I. G., Teslyuk, V. M., Opotiak, Yu. V., & Oliinyk, O. O. (2023). 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. https://doi.org/10.23939/ujit2023.02.097
15. Torskyi, O. I., & Hrytsiuk, Y. I. (2025). Application of machine learning to enhance the efficiency of automated software testing. *Scientific Bulletin of UNFU*, 35(4), 142–149. https://doi.org/10.36930/40350416
16. Wang, X., Yang, J., & Darrell, T. (2024). Segment Anything without Supervision. arXiv:2406.20081v1. https://doi.org/10.48550/arXiv.2406.20081
17. Yang, H., Yin, H., Shen, M., Molchanov, P., Li, H., & Kautz, J. (2023). Global Vision Transformer Pruning with Hessian-Aware Saliency. arXiv:2110.04869v2. https://doi.org/10.48550/arXiv.2110.04869
18. Zhang, C., Han, D., Qiao, Y., Kim, J. U., Bae, S., Lee, S., & Hong, C. S. (2023). Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. arXiv:2306.14289v2. https://doi.org/10.48550/arXiv.2306.14289
19. Zhang, C., Han, D., Zheng, S., Choi, J., Kim, T., & Hong, C. S. (2023). MobileSAMv2: Faster Segment Anything to Everything. arXiv:2312.09579v1. https://doi.org/10.48550/arXiv.2312.09579
20. Zhang, Y., Konz, N., Kramer, K., & Mazurowski, M. A. (2025). Quantifying the Limits of Segmentation Foundation Models: Modeling Challenges in Segmenting Tree-Like and Low-Contrast Objects. arXiv:2412.04243v3. https://doi.org/10.48550/arXiv.2412.04243
21. Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., & Wang, J. (2023). Fast Segment Anything. arXiv:2306.12156v1. https://doi.org/10.48550/arXiv.2306.12156

Б. П. Борківський, В. М. Теслюк

Національний Університет "Львівська політехніка", м. Львів, Україна

СПЕЦІАЛІЗОВАНЕ СТРУКТУРНЕ СПРОЩЕННЯ БАЗОВИХ МОДЕЛЕЙ ДЛЯ СИСТЕМ ТЕХНІЧНОГО ЗОРУ РОБОТІВ У ПРИМІЩЕННЯХ

Проаналізовано проблему впровадження масштабних базових моделей комп'ютерного зору в робото-технічні системи реального часу, що наразі істотно обмежено значними обчислювальними витратами та затримками під час інференсу. Проаналізовано обмеження методів спрощення структури сегментаційних моделей загального призначення, які часто не забезпечують збереження семантичної точності та деталізації масок під час перенесення моделей у спеціалізовані середовища, такі як внутрішня навігація автономних роботів. Представлено фреймворк доменно-специфічної структурної оптимізації внутрішньої структури сегментаційної моделі, який розроблений для трансформації важковагової архітектури моделі сегментації зображень у високоточний інструмент, адаптований для виконання спеціалізованих завдань візуального сприйняття. Досліджено значення багатоетапного конвеєра структурного спрощення моделі сегментації зображень, починаючи із критичної фази адаптації сегментаційної моделі, під час якої повномасштабна модель-вчитель спеціалізується на формуванні стійких апріорних знань про цільові текстури об'єктів та геометрію приміщень, що створює доменно-орієнтований базовий рівень для передачі знань. Реалізовано алгоритм почергового спрощення структури моделі сегментації зображень, який здійснює декомпозицію енкодера на незалежні внутрішні вимірності, що дає змогу здійснювати послідовне структурне спрощення архітектури сегментаційної моделі зі збереженням архітектурної цілісності цієї моделі для узгодження ознак на проміжних рівнях. Інтегровано надійний конвеєр аугментації даних на етапах дистиляції та відновлення сегментаційної моделі, впроваджуючи складні геометричні та фотометричні трансформації зображень для стабілізації процесу навчання сегментаційної моделі та мінімізації ризику її перенавчання за умов дефіциту розмічених даних. Показано, що такий комплексний підхід до спрощення внутрішньої структури сегментаційної моделі забезпечує скорочення загальної кількості навчальних параметрів на 73,3 % та зменшення кількості операцій множення з накопиченням (MACs) на 74,3 %, що фактично дає можливість подвоїти швидкість інференсу – від 7 до 15 кадрів за секунду (FPS). Визначено, що оптимізована модель-учень демонструє тільки незначне зниження точності (приблизно на 1 %) порівняно зі спеціалізованою моделлю-вчителем, стабільно перевершуючи показники стандартного алгоритму SlimSAM як за деталізацією семантичних масок, так і за загальною стійкістю до факторів зовнішнього середовища. Встановлено, що розроблена модель сегментації зображень виявляє значно вищу резистентність до варіацій освітлення та впливу тіней, успішно вирішуючи критичну проблему базових архітектур сегментації зображень, таких як FastSAM, які часто спотворюють маски сегментації за умов висококонтрастного внутрішнього освітлення. З'ясовано, що інтеграція доменно-специфічного спрощення та дистиляції з аугментацією даних забезпечує дещо ефективніший та стабільніший шлях для впровадження складних систем інтелектуального зору реального часу на автономних платформах з обмеженими обчислювальними ресурсами.

Ключові слова: оброблення зображень; оброблення даних; сегментація зображень; тренування моделей; сегментаційні моделі загального призначення.