Optimal Prompting in SAM for Few-Shot and Weakly Supervised Medical Image Segmentation

Lara Siblini[1,2], Gustavo Andrade-Miranda[1](✉)[0000−0002−6499−5655], Kamilia Taguelmimt[1], Dimitris Visvikis[1], and Julien Bert[1]

[1] LaTIM, UMR1101, INSERM, University of Brest, France
andradema@univ-brest.fr
[2] Université de Bourgogne, Dijon, France

Abstract. Recent advancements in medical image segmentation have been driven by deep learning's capability to extract rich features from extensive datasets. However, these improvements rely heavily on large annotated datasets, which pose significant challenges in the resource-intensive medical field. Foundational models, such as Meta's Segment Anything Model (SAM), have been developed to address these challenges. SAM has demonstrated exceptional zero-shot performance, often rivaling or surpassing fully supervised models across various tasks. Nonetheless, SAM cannot be directly applied to medical image segmentation due to domain shift, making it necessary to fine-tune the model using prompts. Reducing the annotation workload is crucial to alleviate the burden and constraints associated with extensive data annotation in the medical field. This study investigates prompt-guided strategies in SAM for medical image segmentation under few-shot and weakly supervised scenarios. We assess various strategies—bounding boxes, positive points, negative points, and their combinations—using two publicly available datasets. Optimal results are achieved using positive-negative points, demonstrating that the SAM model can perform comparably to established methods in hepatic vascular and prostate cancer segmentation, even with minimal examples. This research aims to advance medical image segmentation by decreasing reliance on extensive annotated data, providing insights into effective prompt utilization, and showcasing SAM's adaptability in specialized medical contexts.

Keywords: Foundational models · prompt tuning · SAM · segmentation

1 Introduction

Recent advancements in medical image segmentation have been driven by significant progress in deep learning techniques [3]. The remarkable potential of deep learning lies in its ability to automatically learn features from extensive datasets, leading to substantial improvements in performance [1]. However, these advancements come at the cost of requiring large quantities of annotated data to achieve optimal results [15]. This dependency on large, annotated datasets poses

L. Siblini and G. Andrade-Miranda—Authors contributed equally.

challenges, particularly in the medical field where acquiring and annotating high-quality data is time-consuming and resource-intensive [20]. Furthermore, conventional segmentation models are often trained within task-specific frameworks, which may not always have access to extensive datasets. This scarcity of data is reinforced by a lack of inter-center variability, restricting the models' ability to generalize to unseen data from different institutions [18]. As a result, these models frequently need to be retrained from scratch for each new application or dataset, further increasing the burden on resources and time.

To overcome the limitations of traditional training methods, many studies have employed transfer learning from pre-trained data on ImageNet [13]. However, this approach often resulted in suboptimal performance when applied to the medical domain [16]. In response, self-supervised learning [2] has emerged as a new paradigm enabling models to learn accurate and meaningful representations of input data without labels, thereby facilitating the development of large-scale foundational models. These comprehensive models, trained on extensive corpora, represent a significant breakthrough in the AI community. They can be easily adapted to task-specific problems through fine-tuning and prompting strategies [4, 6], significantly reducing the reliance on extensive domain-specific training data. One of the most notable examples of these models is the Segment Anything Model (SAM) from Meta [12], which was trained on an extensive corpus comprising 1 billion annotations across 11 million images.

SAM can segment objects using various human input prompts like dots, bounding boxes, or text. Evaluations highlight its exceptional zero-shot performance, often matching or surpassing fully supervised models across diverse tasks [19]. With these strengths, SAM holds promise for various vision applications, including medical image segmentation. However, due to the substantial domain shift between natural and medical images, directly applying SAM is impractical [8]. Therefore, fine-tuning the SAM model with optimal prompting strategies is crucial for achieving accurate region segmentation. Previous studies have evaluated the zero-shot performance of SAM [14]. However, our focus is on fine-tuning SAM using few-shot learning and optimized prompting strategies. This approach aims to alleviate the burden of extensive manual labeling by leveraging weakly supervised segmentation with minimal annotated data through points or box prompts. This method enhances the efficiency of annotating large datasets and improves segmentation accuracy.

In this study, we explore prompt-guided strategies within the SAM model for medical image segmentation under few-shot and weakly supervised learning scenarios. Our primary contributions can be summarized as follows:

- We conduct an extensive assessment to evaluate various prompting strategies in the context of few-shot learning, with the aim of enhancing medical segmentation performance for hepatic vascular and prostate cancer.
- We compare the performance of the best prompting strategy with that of fully supervised state-of-the-art methods trained on limited data. This comparison involves two transformer-based segmentation networks and the well-known nnU-Net framework [11].

– We provide insights into the optimal location, position, and number of points required to enhance segmentation performance. Additionally, we discuss the potential benefits and challenges associated with each prompting method, offering a comprehensive evaluation of their efficacy in hepatic vascular and prostate cancer segmentation.

2 Methods

2.1 Overview of SAM

SAM represents the largest foundational model for natural image segmentation, trained on the extensive SA-1B dataset. It is designed to handle various user prompts for image segmentation. SAM consists of three core components: an Image Encoder (IE) based on the Vision Transformer (ViT) architecture [5], which extracts features from images; a Prompt Encoder (PE) that processes different types of prompts, including points, bounding boxes, masks, and text; and a lightweight Decoder (MD) that translates the combined image and prompt embeddings into segmentation results.

2.2 Prompting Strategies

In the following subsections, we will explore various prompting strategies designed to guide the foundational SAM model in medical image segmentation. These strategies include the use of bounding boxes, positive points, simultaneous positive and negative points, and a hybrid approach that integrates both bounding boxes and points. Each strategy provides different levels of guidance to the model, aiming to enhance segmentation accuracy while reducing reliance on extensive annotated datasets.

Box prompting Box prompting uses bounding boxes around regions of interest to train the model, providing a weakly supervised approach to guide the segmentation process. These bounding box coordinates are generated based on the ground truth segmentation. To emulate annotator variability, the bounding boxes are randomly expanded by 5 to 20 pixels around the ground truth. This expansion extends the bounding box slightly beyond the non-zero elements, introducing a form of human variability in the bounding box dimensions.

Point prompting This strategy involves using labeled points to guide the model in the segmentation process. Each point is associated with a label: positive points, situated inside the region of interest (ROI), are labeled as +1 (depicted as red dots in Figure 1), while negative points, located outside the ROI and corresponding to the background, are labeled as -1 (depicted as green dots in Figure 1). Generally, we can differentiate between two strategies: utilizing solely positive points or integrating both positive and negative points.

Positive points prompting: This method involves placing one or several points within the target regions. In our experiments, we tested both strategies. In the first strategy, each disjoint region is represented by a single point positioned near the centroid of the object. To reflect annotator variability, the points are placed randomly within a circle of radius r, where r is smaller than the radius of the largest inscribed circle. The first column of Figure 1

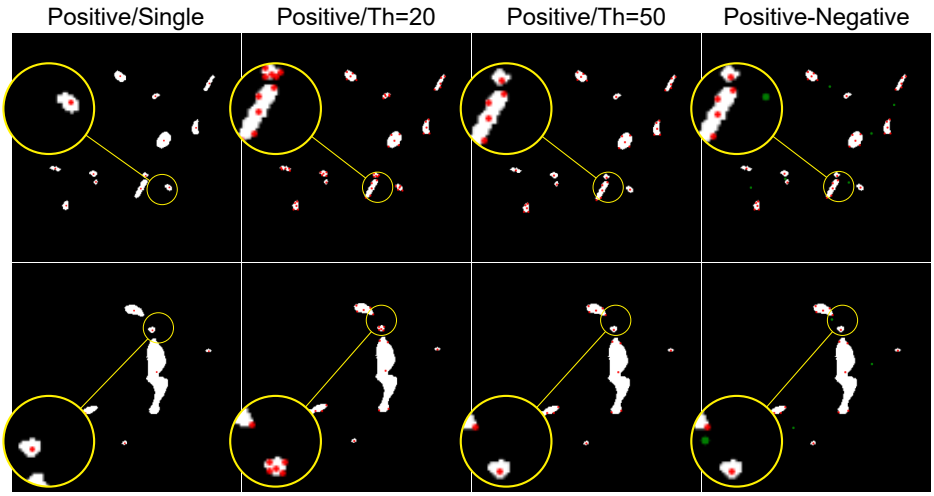

Fig. 1. Illustration of different point prompting tuning strategies. The first column represents single positive points. The second and third columns depict multiple positive points with different Th. The last column represents the positive-negative strategy with a threshold of $Th = 50$. Positive points are represented in red, while negative points are shown in green.

illustrates two examples of single-point prompting, with each positive point randomly positioned near the object's centroid to simulate realistic annotation conditions.

To determine the use of single or multiple points per region, we base our decision on the area of each object to be segmented. We evaluate a series of thresholds Th and decide whether to include additional points based on these values. When an object's area exceeds Th, additional points are randomly distributed along the contours with a margin of error relative to the ground truth. This method enhances segmentation in larger regions and those with varying contrast, thereby avoiding under-segmentation problems. The second and third columns in Figure 1 illustrate this effect for threshold values of $Th = 20$ and $Th = 50$. The smaller region in the second column contains more points compared to the same region in the third column, where $Th = 50$. However, in both cases, the larger region consistently contains more than one point.

Positive-negative points prompting: This method works similarly to the previous one, with the distinction of also including negative points. Positive points are assigned based on the same threshold principle, while negative points are strategically placed between the contours of nearly disjoint regions, where the model is prone to misinterpreting them as a single region. The objective is to mitigate over-segmentation issues that arise due to the presence of closely situated objects. The last column of Figure 1 illustrates

this concept, showing a negative point (green) placed between two closely situated objects.

Hybrid - Box-Points prompting In the final implementation, a combination of bounding boxes and positive-negative points was used. This strategy incorporates the previously described method of assigning multiple positive and negative points along with bounding boxes. This hybrid method aims to leverage the strengths of both point-based and bounding box techniques to achieve more precise and reliable segmentation results.

3 Experiments

3.1 Datasets

In our experiments, we employ the `3D-IRCADb` [7] and `PICAI` [17] datasets to evaluate the SAM model. For both datasets, we perform evaluations in the context of few-shot learning, focusing on scenarios with a limited number of annotated samples.

3D-IRCADb: This dataset comprises 20 contrast-enhanced CT volumes with varying image resolutions and vessel structures, featuring hepatic tumors in 75% of the cases. For this study, images were resized axially using the liver bounding box in each CT scan to a resolution of 256×256 with 128 slices per patient. The original anisotropic voxel spacing was standardized to an isotropic spacing of 1×1×1 mm.

PICAI: This dataset, obtained from the corresponding grand challenge, comprises over 1500 publicly available cases. Each case features three imaging modalities: T2-weighted imaging (T2W), diffusion-weighted imaging (DWI), and apparent diffusion coefficient maps (ADC). Specifically, we examine 40 cases where clinically significant prostate cancer (csPCa) is identified.

3.2 Implementation details

We experimented with various prompting strategies, including bounding boxes, single and multiple positive points, positive-negative points, and combinations of boxes and points. For all configurations, we fine-tuned only the mask decoder, keeping the image and prompt encoder frozen. All experiments were conducted using the ViT-B model for inference and fine-tuning, treating the final segmentation task as binary for both datasets. Models were fine-tuned for 200 epochs using the Adam optimizer with an initial learning rate of 1e-5, and other hyperparameters set according to the original SAM paper. Training was conducted with a batch size of 10 on a single GPU with 48GB of memory. For the `3D-IRCADb` dataset, we trained and validated on 18 volumes and tested on 2 volumes. For the `PICAI` dataset, we used 20 volumes for training and validation, and an additional 20 for testing.

3.3 Evaluation

We first compare various prompting strategies with three state-of-the-art (SOTA) models: Swin-UNETR [9], UNETR [10], and nnU-Net [11]. This evaluation is conducted on the `3D-IRCADb` dataset, assessing Dice, Jaccard, and

Precision metrics. All models are trained using the same number of volumes. After identifying the optimal prompting strategy, we evaluate its generalizability by applying it to a new downstream task: prostate tumor segmentation using the `PICAI` dataset. For this assessment, we employ the Dice score as the evaluation metric. To further evaluate the few-shot capabilities of the optimal prompting strategy, we fine-tuned the model on varying numbers of training cases (ranging from 6 to 16). Additionally, we present qualitative results for both the `PICAI` and `3D-IRCADb` datasets, visually comparing the performance of the prompting strategies and state-of-the-art methods. This comparison aims to highlight the strengths of the prompting approach and demonstrate the effectiveness of the optimal strategy across different scenarios.

4 Results and discussion

Initially, we evaluated the performance of various prompting strategies on the `3D-IRCADb` dataset using two test volumes (see Table 1). The best performance was achieved with Positive-Negative prompting, yielding Dice scores of 0.74 and 0.70. Close to this strategy were the hybrid and multiple positive points approaches, with Dice scores of 0.72 and 0.68, respectively. Among the state-of-the-art (SOTA) methods, the best-performing model was nnU-Net, achieving Dice scores of 0.74 and 0.60. It is important to note that all prompting strategies and supervised methods were trained using the same amount of data (18 samples). Particularly noteworthy is that the worst-performing prompting strategy was the one based on bounding boxes. This may be due to the inclusion of false positive pixels within the bounding box, which can negatively impact the segmentation task.

Prompting	Dice (↑)		Jaccard (↑)		Precision (↑)	
	S1	S2	S1	S2	S1	S2
Box	0.54	0.48	0.40	0.33	0.71	0.61
Positive point/Single	0.67	0.62	0.52	0.46	0.76	0.69
Positive point/Th=50	0.70	0.64	0.55	0.50	0.82	0.71
Positive point/Th=20	0.72	0.67	0.57	0.53	0.83	0.71
Positive-Negative/Th=20	**0.74**	**0.70**	**0.60**	**0.55**	0.83	0.74
Hybrid	0.72	0.68	0.58	0.53	0.81	0.71
UNETR	0.51	0.43	0.34	0.28	0.56	0.65
nnU-Net	**0.74**	0.60	0.59	0.43	**0.90**	**0.81**
Swin-UNETR	0.63	0.51	0.46	0.34	0.76	0.74

Table 1. Results of the different prompting strategies and SOTA methods for the two test volumes in the `3D-IRCADb` dataset.

We also investigated whether the use of a positive-negative points prompting, maintains good performance in a different downstream task. We then compared this strategy against SOTA models, with the results presented in Figure 2. It is evident that SAM significantly outperforms the SOTA models, with all predictions achieving Dice scores of at least 0.6. These results underscore the impor-

tance of foundational models and the effectiveness of prompt tuning, especially in data-limited scenarios.

To evaluate the few-shot capabilities of the positive-negative prompting strategy, we trained SAM varying the numbers of training cases. The results in Figure 3 show that the positive-negative fine-tuning outperformed supervised methods even with only six training cases. The only exception is in $S1$, where nnU-Net performs better. These findings highlight the robust few-shot learning capability of the SAM model. Performance improvements became significant after using more than 10 training cases. However, further investigation is needed to determine the plateau point beyond which additional training samples do not yield substantial gains.

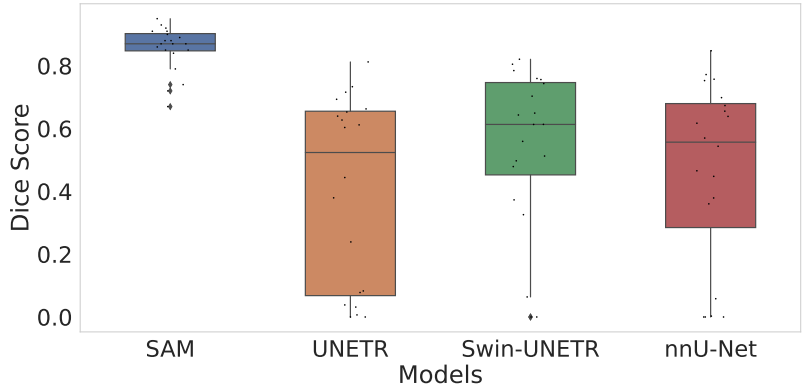

Fig. 2. Boxplots comparing the Positive-Negative prompting strategy with UNETR, Swin-UNETR, and nnU-Net models for `PICAI` datasets.

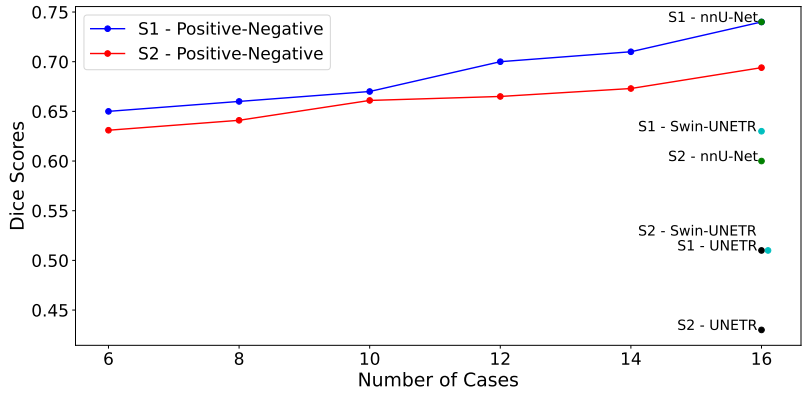

Fig. 3. Illustration of Dice scores on the `3D-IRCADb` dataset using the positive-negative prompting strategy. Lines represent performance with varying training sample sizes (6 to 16). Colored dots indicate Dice scores of SOTA methods for comparison.

Figure 4 illustrates the difference between using only positive points and using both positive and negative points. For instance, the inclusion of negative points helps to accurately predict small regions that disappear when only positive points

are used (first and second columns). Additionally, the absence of negative points (the last two columns) leads to small regions being segmented as a single entity, resulting in over-segmentation. In the prostate cancer segmentation task depicted in Figure 5, although the models sometimes struggle to accurately delineate the shape of the tumor, they consistently detect its presence.

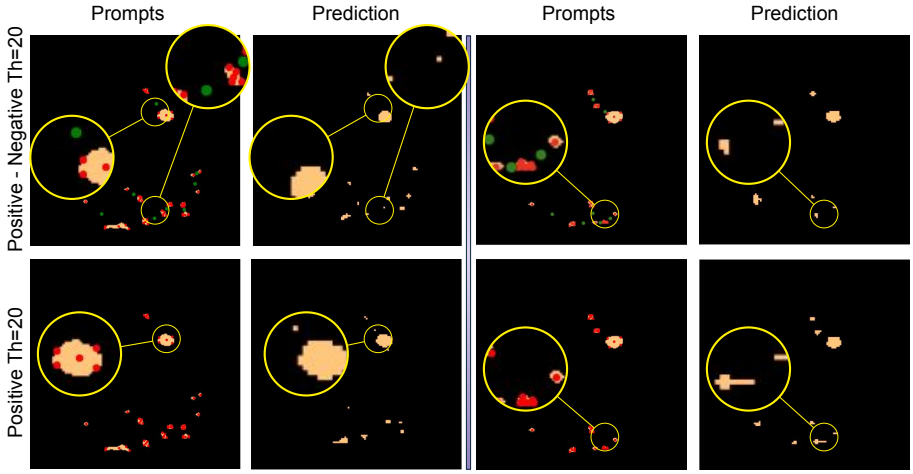

Fig. 4. Visual comparison of positive-negative versus positive points prompting. Magnified areas are highlighted in yellow, with positive points shown in red and negative points in green.

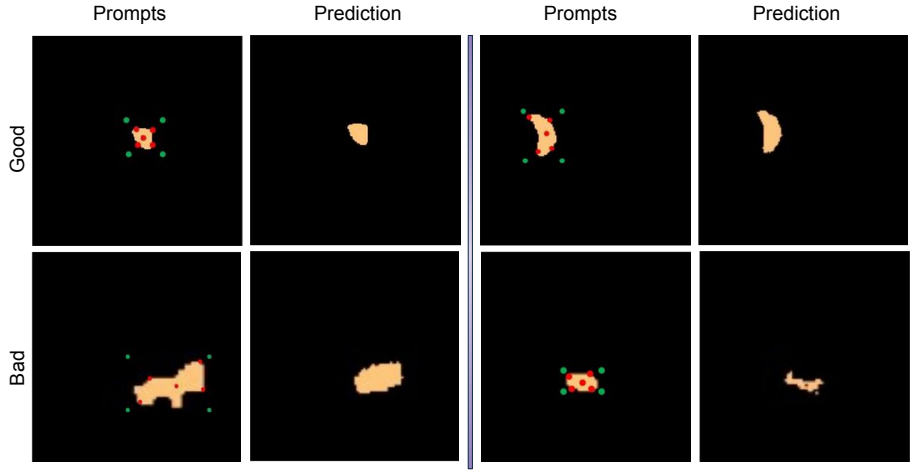

Fig. 5. Thumbnails of good and poor Segmentation results using positive-negative prompts on the `PICAI` Dataset

5 Conclusion

By employing various prompting fine-tuning strategies on the SAM model, we have demonstrated its versatility across two medical image segmentation tasks.

Our study reveals that incorporating both positive and negative points is the most effective approach for enhancing segmentation performance. Specifically, our findings indicate that multiple points may be necessary depending on the size of the object, particularly near its boundary. The inclusion of negative points also helps to mitigate over-segmentation and under-segmentation issues, especially when disjoint objects are in close proximity. This work provides valuable insights and guidelines for the optimal placement of prompts in the context of medical imaging, thereby reducing the reliance on fully annotated datasets.

References

1. Andrade-Miranda, G., Jaouen, V., Tankyevych, O., Cheze Le Rest, C., Visvikis, D., Conze, P.H.: Multi-modal medical transformers: A meta-analysis for medical image segmentation in oncology. Computerized Medical Imaging and Graphics **110**, 102308 (2023)
2. Balestriero, R., Ibrahim, M., Sobal, V., Morcos, A., Shekhar, S., Goldstein, T., Bordes, F., Bardes, A., Mialon, G., Tian, Y., Schwarzschild, A., Wilson, A.G., Geiping, J., Garrido, Q., Fernandez, P., Bar, A., Pirsiavash, H., LeCun, Y., Goldblum, M.: A cookbook of self-supervised learning (2023)
3. Conze, P.H., Andrade-Miranda, G., Singh, V.K., Jaouen, V., Visvikis, D.: Current and emerging trends in medical image segmentation with deep learning. IEEE Transactions on Radiation and Plasma Medical Sciences **7**(6), 545–569 (2023)
4. Davila, A., Colan, J., Hasegawa, Y.: Comparison of fine-tuning strategies for transfer learning in medical image classification. Image and Vision Computing **146**, 105012 (2024)
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.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
6. Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S.A., Hospedales, T.: Parameter-efficient fine-tuning for medical image analysis: The missed opportunity. In: Medical Imaging with Deep Learning (2024)
7. Garret, G., Vacavant, A., Frindel, C.: Deep vessel segmentation based on a new combination of vesselness filters. ArXiv **abs/2402.14509** (2024)
8. Gu, H., Dong, H., Yang, J., Mazurowski, M.A.: How to build the best medical image segmentation algorithm using foundation models: a comprehensive empirical study with segment anything model (2024)
9. Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H., Xu, D.: Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (2022)
10. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 272–284 (2022)
11. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2020)
12. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)

13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems. vol. 25. Curran Associates, Inc. (2012)
14. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: An experimental study. Medical Image Analysis **89**, 102918 (2023)
15. Moor, M., Banerjee, O., Abad, Z.S.H., et al.: Foundation models for generalist medical artificial intelligence. Nature **616**, 259–265 (2023)
16. Raghu, M., Zhang, C., Kleinberg, J., Bengio, S.: Transfusion: Understanding transfer learning for medical imaging. Advances in neural information processing systems **32** (2019)
17. Saha, A., Bosma, J., Twilt, J., van Ginneken, B., Yakar, D., Elschot, M., Veltman, J., Fütterer, J., de Rooij, M., henkjan huisman: Artificial intelligence and radiologists at prostate cancer detection in MRI — the PI-CAI challenge. In: Medical Imaging with Deep Learning, short paper track (2023)
18. Sallé, G., Andrade-Miranda, G., Conze, P.H., Boussion, N., Bert, J., Visvikis, D., Jaouen, V.: Cross-modal tumor segmentation using generative blending augmentation and self-training. IEEE Transactions on Biomedical Engineering pp. 1–12 (2024)
19. Wald, T., Roy, S., Koehler, G., Disch, N., Rokuss, M.R., Holzschuh, J., Zimmerer, D., Maier-Hein, K.: Sam.md: Zero-shot medical image segmentation capabilities of the segment anything model. In: Medical Imaging with Deep Learning, short paper track (2023)
20. Yang, F., Zamzmi, G., Angara, S., Rajaraman, S., Aquilina, A., Xue, Z., Jaeger, S., Papagiannakis, E., Antani, S.K.: Assessing inter-annotator agreement for medical image segmentation. IEEE Access **11**, 21300–21312 (2023), epub 2023 Feb 27