FaM-SAM's Lightning-Fast Precision for Enhanced Diagnostics

Pushkar Ambastha

Abstract

The recently proposed segment anything model (SAM) has significantly influenced many computer vision tasks. It is becoming a foundation step for many high-level tasks, like image segmentation, image captioning, and image editing. However, its substantial computation costs prevent it from broader applications in industry scenarios (especially medical frameworks). The computation mainly comes from the Transformer architecture at high-resolution inputs. This paper proposes a speed-up alternative method for this fundamental task with comparable performance. By reformulating the task as segment generation and prompting, we find that a regular CNN detector with an instance segmentation branch can also accomplish this task. Also, directly applying SAM to medical image segmentation cannot perform satisfactorily on multi-modal and multi-target medical datasets. Many insights are drawn to guide future research to develop foundation models for medical image analysis. We take inferences from these models on selected public datasets and find patterns to extrapolate to create a new model proposition.

1. Introduction

The main ideas revolve around the fact that the performance of pre-existing models could be more helpful against complex datasets in healthcare. Deep learning-based models have shown great promise in medical image segmentation because they can learn complex image features and provide accurate and efficient segmentation results. However, current models are often tailored to specific imaging modalities and targets, and their generalization ability could be improved. Therefore, developing foundation models that can adapt to various medical imaging modalities and segmentation targets is paramount in advancing medical image analysis. These models require excessive fine-tuning on medical datasets, and the inference time must be reduced. We should measure the tradeoff between the accuracy and throughput of these models. The recent studies allow us to explore and best use diverse existing work. The work extrapolates the latency performance of FastSAM and the trained weights of MedSAM by analyzing the performance of selected datasets. Our method aims to achieve better accuracy and inference time for deployment in the healthcare sector. We test the proposed model on various medical dataset that challenges previous foundation model like SAM. (discussed with examples).

2. Related work

2.1. SAM: generalization and versatility

SAM diverges from traditional segmentation frameworks by introducing a novel promptable segmentation task supported by a flexible prompting-enabled model architecture and vast and diverse sources of training data. A data engine was proposed to build a cyclical process that utilizes the model to facilitate data collection and leverages the newly collected data to enhance the model's performance. SAM utilizes a transformer-based architecture, which is highly effective in natural language processing and image recognition tasks. Specifically, SAM uses a vision transformer-based image encoder to extract image features and compute an image embedding and a prompt encoder to embed prompts and incorporate user interactions. Then removed information from two encoders is combined into a lightweight mask decoder to generate segmentation results based on the image embedding, prompt embedding, and output token.

Image Encoder. The vision transformer in the image encoder is pre-trained with a masked auto-encoder, which is minimally adapted to process high-resolution (i.e., 1024×1024) images. After the image encoder, the obtained image embedding is 16× downscaled to 64×64.

Prompt Encoder. Two prompts are considered for prompt encoders: sparse (points, boxes, text) and dense (masks). SAM

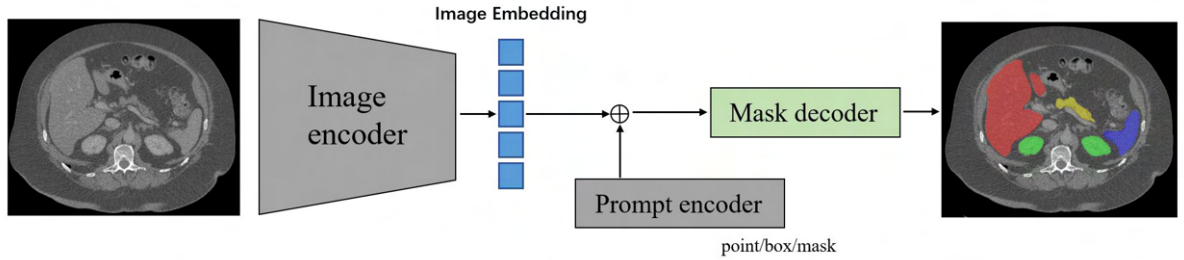

Figure 1. Overview of the architecture of Segment Anything Model (SAM).

employs positional encoding combined with learned embeddings to represent points and boxes. Specifically, attributes are encoded by two learnable tokens for specifying foreground and background, and the bounding box is encoded by the point encoding of its top-left corner and bottom-right corner. The pre-trained text encoder from CLIP encodes the free-form text. Dense mask prompts with the exact spatial resolution as the input image are embedded using convolutions and summed element-wise with the image embedding.

Mask Decoder. The mask decoder is characterized by its lightweight design, which consists of two transformer layers with a dynamic mask prediction head and an Intersection-over-Union (IoU) score regression head. The mask prediction head can produce three 4× downscaled masks, corresponding to the whole object, part, and subpart of the object, respectively. The linear focal and Dice loss combination supervised the output prediction during training. A data engine is built for label-efficient training—specifically, professional annotators first label masks through interactive segmentation. Then, less prominent objects which are ignored in the predictions of SAM will be labeled manually. Finally, a fully automatic stage is conducted, in which confident and stable pseudo masks are selected as annotations.

2.2. Med-SAM: Extending the Usability of SAM to Medical Images

Ma et al. introduce MedSAM for universal image segmentation by curating a diverse and comprehensive medical image dataset containing over 200,000 masks with 11 modalities and developing fine-tuning approach to adapt SAM to medical image segmentation. However, the overall performance is still behind specialist models for medical image segmentation.

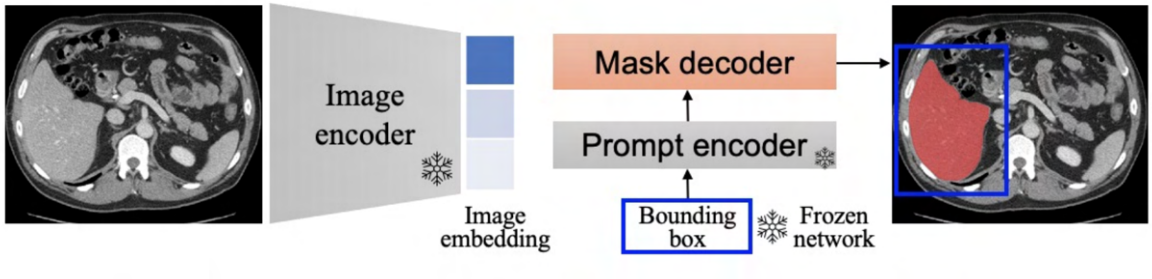

Figure 2. MedSAM: Fine-tuning Segment Anything Model for medical image segmentation by freezing the image encoder and prompt encoder and only fine-tuning the mask decoder.

2.3. Fast-SAM: lightweight and efficient

The proposed FastSAM achieves a comparable performance with SAM and runs 50x faster than SAM (32×32) and 170x faster than SAM (64×64). The running speed makes it a good choice for industrial applications, such as road obstacle

detection, video instance tracking, and image manipulation. The authors rethink the segment of anything task and the model architecture choosing and proposing an alternative solution with 50 times faster running speed than SAM-ViT-H (32×32). The experiments show that FastSAM can solve multiple downstream tasks well.

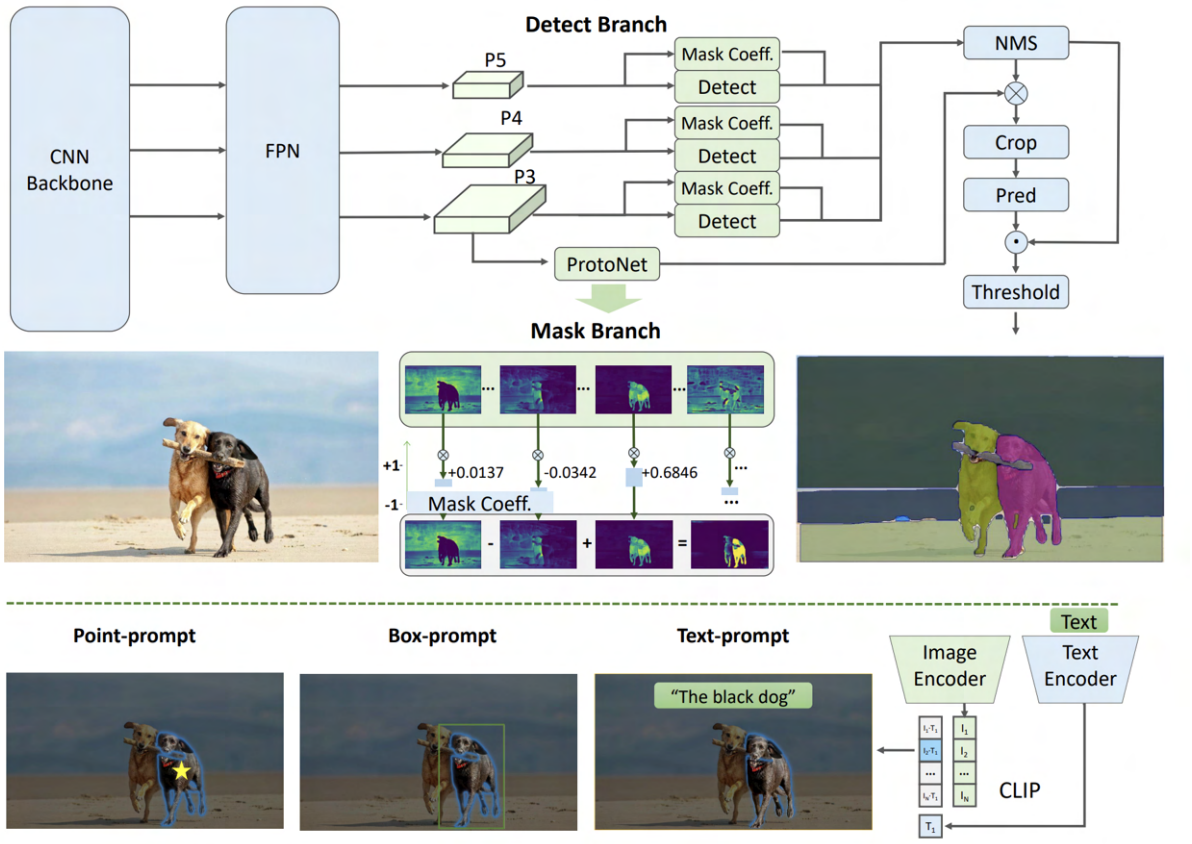

Figure 3. The framework of FastSAM. It contains two stages: All-instance Segmentation (AIS) and Prompt-guided Selection (PGS). We use YOLOv8-seg to segment all objects or regions in an image. Then we use various prompts to identify the specific object(s) of interest. It mainly involves the utilization of point prompts, box prompts, and text prompts. The text prompt is based on CLIP.

2.4. How SAM performs on Medical Image Segmentation?

We review recent studies to benchmark SAM on different medical image segmentation tasks compared to domain-specific segmentation methods. Generally, SAM requires substantial human information to obtain overall moderate segmentation performance using only a few points or bounding box prompts. Overall, these evaluation results on different datasets show that SAM had limited generalization ability when directly applied to medical image segmentation, which varies significantly across other datasets and tasks. SAM delivers remarkable performance comparable to state-of-the-art methods in some specific objects and modalities. However, SAM is imperfect or even fails in more challenging situations when the segmentation targets have weak boundaries with low-contrast and more minor and irregular shapes, which aligns with other investigations. In most cases, the subpar segmentation performance of SAM is not sufficient and satisfying for further applications, especially for medical image segmentation tasks where extremely high accuracy is demanded. Since SAM is pre-trained on the SA-1B dataset consisting of natural images, the objects usually have robust edge information, significantly different from medical images. Therefore, directly applying SAM to these unseen and challenging medical image segmentation tasks may need more performance.

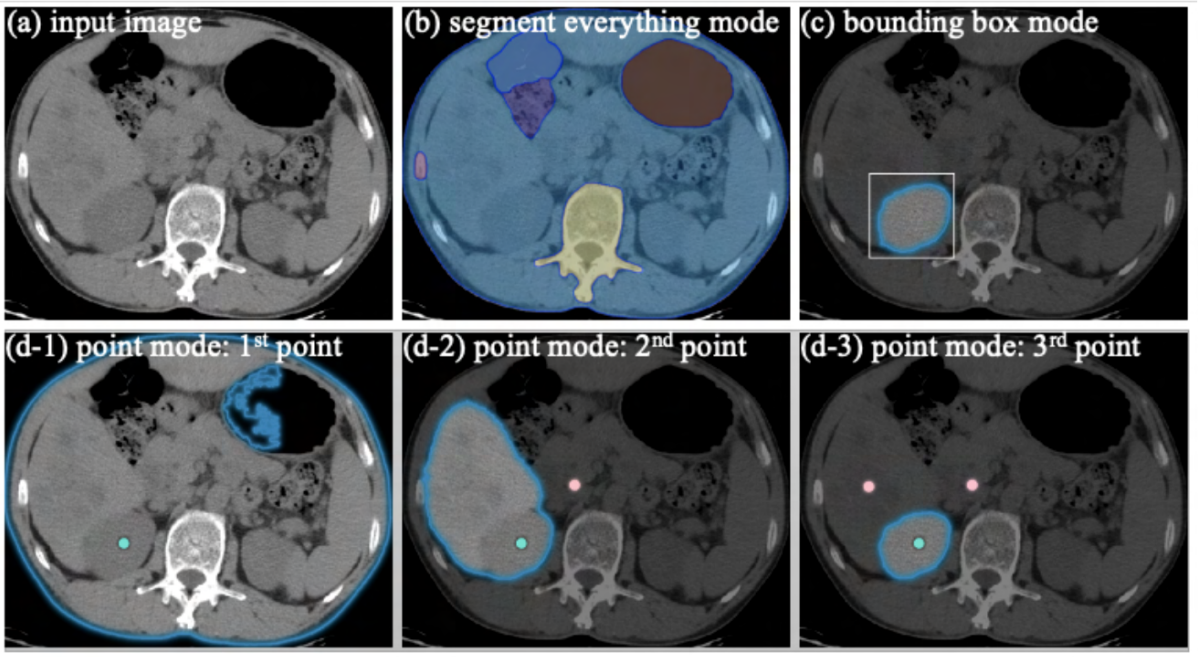

Figure 4. Segmentation results of SAM based on different segmentation modes.

3. Methodology

3.1. Background on SAM

Here, we first summarize the structure of SAM and how it works. SAM consists of a ViT-based image encoder and a prompt-guided mask decoder. The image encoder takes the image as the input and generates an embedding, then fed to the mask decoder. The mask decoder generates a mask to cut out any object from the background based on a prompt like a point (or box). Moreover, SAM allows generating multiple masks for the same prompt to address the ambiguity issue, which provides valuable flexibility. Considering this, this work maintains the pipeline of SAM first to adopt a ViT-based encoder to generate image embedding and then to adopt a prompt-guided decoder to generate the desired mask. This pipeline is optimally designed for the "segment anything," which can be used for the downstream task of "segment everything."

3.2. Project goal

This project aims to generate a deployment-friendly SAM (FaM-SAM) that achieves satisfactory performance in a lightweight manner and is much faster than the original SAM. The prompt-guided mask decoder in the original SAM has less than 4M parameters and is thus considered light. Given an image embedding processed by the encoder, as shown in their public demo, SAM can work in resource-constrained devices since the mask decoder is lightweight. However, the default image encoder in the original SAM is based on ViT-H with more than 600M parameters, which is very heavyweight and makes the whole SAM pipeline incompatible with common usage devices. Therefore, the key to obtaining a usage-friendly SAM lies in replacing the heavyweight image encoder with a lightweight one, which also automatically keeps all its functions and characteristics of the original SAM. In the following, we elaborate on our proposed method for achieving this project goal. This project also seeks to analyze the performance of SAM and Fast-SAM on selected medical datasets (e.g., Covis-19 CT-Scans, Cancer and Influenza Image datasets). The idea also involves the combination of these two models - Fast-SAM and MedSAM over the selected datasets to extrapolate the results to create a new model called Fast Medical SAM (FaM-SAM).

3.3. Proposed Method

The method combines Fast-SAM and Med-SAM, where the model is trained on medical datasets. Like Fast-SAM, we segment all objects or regions in an image using YOLOv8 (finetuned on the medical dataset) and then use prompts to a

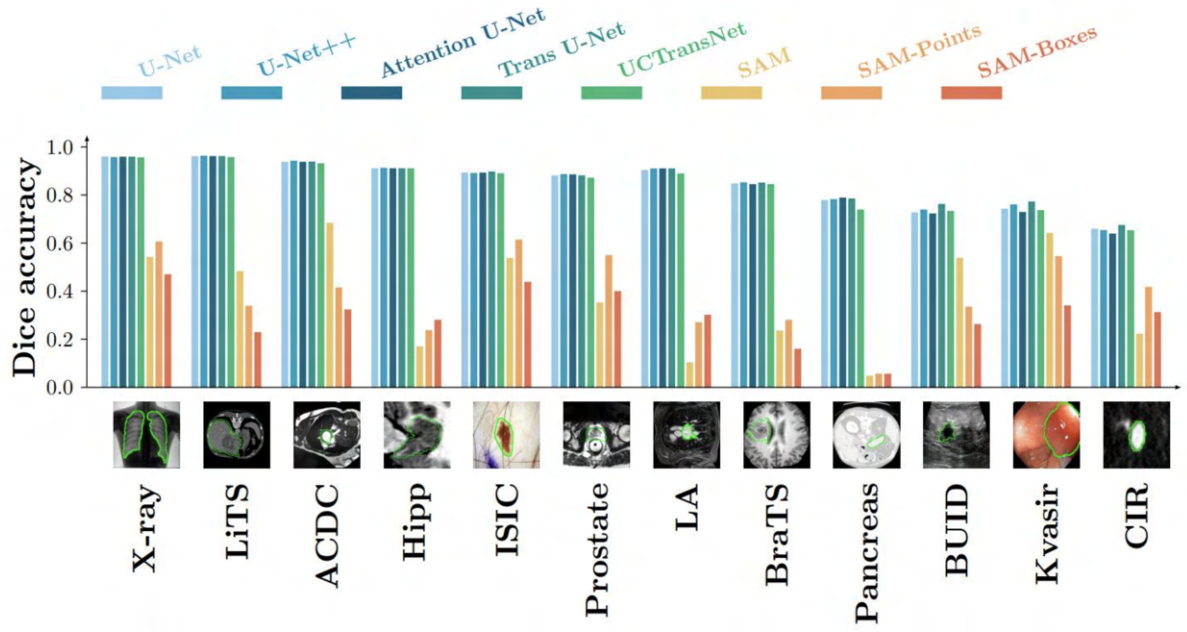

Figure 5. Visual comparison of Dice Similarity Coefficient between SAM and existing segmentation networks on 12 medical image segmentation datasets.

Table 1. Comparison of the parameters and speed for the image encoder in original SAM and Fast-SAM. The inference speed is measured on a single GPU.

	Vanilla SAM	Fast-SAM
Parameters	0.6G	68M
Speed	452ms	80ms

specific object(s) of interest. It mainly involves the utilization of point prompts, box prompts, and text prompts. However, similar to Med-SAM, we keep the image encoder frozen. The prompt encoder encodes the positional information of the bounding box or the text embeddings of the text extracted using the CLIP model. This prompt encoder can be reused, so we freeze that as well. The remaining part that requires fine-tuning is the mask decoder. Extrapolation has been done for the results consisting time performance of Fast-SAM and the accuracy of Med-SAM on medical datasets.

4. Experiments and Results

4.1. Experimental Setup

4.1.1. Lightweight Image Encoder

The Encoder is smaller than the heavyweight image encoder of the original SAM model, and the performance of these two image models is at a trade-off with the accuracy with inference time. The inference speed is checked on a single Tesla P100 GPU for both models.

4.1.2. Training and evaluation details

Similar to MedSAM, we are training the models on selected medical datasets and taking evaluations on the test images. We used Tesla-P100 GPU for training and Inference purposes, and all the datasets are publicly available. All the codes and demos are available here. Here, we use the segment everything mode of the model for the instance segmentation task. More stress has been given to text prompts than the other methods since working with them makes it more uncomplicated than

Table 2. Performance comparison between Fast-SAM and SAM on Medical image segmentation tasks.

Metric	Datasets	Modality	Fast-SAM	Vanilla SAM(GT)
mean IOU	Abdomen Covid-19	CT	0.22	1.0
	Lung Covid-19	X-Ray	0.93	1.0
	CRAG	Pathology	0.27	1.0
	GLAS	Pathology	0.25	1.0
	CPM-15	Pathology	0.45	1.0
	Kumar	Pathology	0.05	1.0
Dice Score Coefficient	Abdomen Covid-19	CT	0.710	1.0
	Lung Covid-19	X-Ray	0.806	1.0
	CRAG	Pathology	0.538	1.0
	GLAS	Pathology	0.703	1.0
	CPM-15	Pathology	0.865	1.0
	Kumar	Pathology	0.150	1.0
Mean Pixel Accuracy	Abdomen Covid-19	CT	0.16	1.0
	Lung Covid-19	X-Ray	0.69	1.0
	CRAG	Pathology	0.29	1.0
	GLAS	Pathology	0.23	1.0
	CPM-15	Pathology	0.49	1.0
	Kumar	Pathology	0.05	1.0

making manual labels using boxes and points. Finally, we compare the performance of SAM, MedSAM, and Fast-SAM on various parameters and extrapolate the time performance of Fast-SAM with the weights of MedSAM as a creation of the proposed model FaM-SAM.

4.2. FaM-SAM performs on par with the original SAM

4.2.1. Ablation study

The Results of these experiments show that with the removal of the components, the models still perform at par under limited resource conditions observed during the deployment of these models. The metrics results show that the heavyweight SAM model outperforms these light models in developing more masks. However, lightweight models like Fast-SAM and FaM-SAM are better at deployment and have various uses in the software industry. We extensively concentrate on the novel approach of including and using the text prompt, as this kind of prompt helps the practitioners (in the medical domain) find the segmented results easily. The Results are a comparison for taking Vanilla SAM as ground truth for verifying Fast-SAM on medical datasets for generating annotations for the unknown image dataset.

4.3. FaM-SAM performs on par with the Med-SAM

The proposed model performs on par with the Med-SAM based on accuracy. The results show that the FaM-SAM performs better than the Med-SAM model regarding inference time. The model is lightweight and efficient in terms of usage, and it can increase the application of such models in diverse medical domains. We show that fine-tuning the masked encoder can significantly improve segmentation tasks and image modalities. However, its performance still needs to catch up to specialist models developed for liver segmentation in abdomen CT images. We observe diverse cases in abdomen tumor and lung COVID-19 infections segmentation, where there is room for improvement, yet comparing performance from the original SAM, performance is far better.

4.4. FaM-SAM outperforms FastSAM

4.4.1. Segment anything v.s. Segment everything

The application for the SAM is seen in the condition where we indent to segment everything in medical industries. We observe that the usage is to create annotations of the images, where these models can perform efficiently. The application can be invoked through the prompts given to the model, like boxes, points, and text. We explore the text in the following content as it becomes helpful in deployment, as it is challenging to draw boxes and points at minute portions of CT-Scans

Table 3. Performance comparison between MedSAM and SAM on Medical image segmentation tasks.MedSAM achieves significant and consistent improvements across all the tasks.

Metric	Dataset	Modality	Med-SAM	SAM	Improve
mean IOU	Breast-Cancer	Ultrasound	0.92	0.452	0.468
	Lung-Covid-19	X-Ray	0.75	0.27	0.48
	Breast Tumor	Ultrasound	0.854	0.780	0.741
Dice Score Coefficient	Breast-Cancer	Ultrasound	0.957	0.452	0.505
	Lung-Covid-19	X-Ray	0.855	0.258	0.597
	Breast Tumor	Ultrasound	0.939	0.803	0.198

Table 4. Comparison between FastSAM and SAM during Inference.

	FastSAM	Vanilla SAM	Ratio
Inference time	84.5ms	452 ms	≈ 5
Preprocess time	9.9ms	72 ms	≈ 7
Post-process time	5.0ms	5.2 ms	≈ 1
Size	68 M	0.6 G	≈ 9

and cross-section images. (confining the discussion to the medical domain only). This can also be easily observed that model architectures like Fast-SAM (YOLO v8) directly generate—mask proposals in a prompt-free manner. To enable promptable segmentation, a mapping algorithm is designed to select the mask from the proposal mask sets. The follow-up works that evaluate its generalization/robustness or investigate its versatility mainly focus on the anything instead of everything mode because the former addresses the foundation task. Therefore, the comparison with FastSAM concentrates primarily on "segment anything," but we also provide a comparison regarding "segment everything" for completeness. The performance of the former is better than the latter, so we propose further results focusing more on it.

4.4.2. FaM-SAM is faster and smaller

The Results on Inference time and size of the model are discussed here. The correct way of comparison is the generation of a single mask proposal from the supplied input image. The Fast-SAM and finetuned versions can generate masks in a prompt-free manner, given their architecture. Still, we analyze the performance in uniform conditions: segment anything mode and text prompts as support.

4.4.3. mIoU comparison under segment anything model

The mIoU comparison has been made for the segment anything model under the resource-limited environment and for the Fast-SAM and SAM models on selected medical datasets. We perform inference and calculate scores considering the ground truth label. This is when we have a previous annotation from other models or manual annotations.

4.4.4. Results for segment everything

The results for the datasets are present as the various metrics and masks are discussed in the tables. We find the metric scores for the selected datasets, comparing them with the ground truth labels where these labels are available. The Results show that these models have room for improvement in these domains. The metric details are available in the Appendix section. The data and codes for these experiments are also available in later areas. These results are replicable and can be used for further studies since the work is based on public datasets.

5. Discussions and Conclusion

The SAM and Fast-SAM propose various novel improvements in the medical industry usage, and the results show that, in terms of results, the models proposed in the paper are at par with the present models and, in the resource constraint environment, perform better since they have lower throughput. The challenges are that the models already fine-tuned, like Med-SAM, have yet to release the datasets they trained the mask-decoder on, so we could only extrapolate and indicate the

Table 5. Performance comparison between Fast-SAM and SAM on Medical image segmentation tasks. Final results and conclusions can be derived from these results.

Metric	Datasets	Modality	Fast-SAM	Vanilla SAM
mean IOU	Abdomen Covid-19	CT	0.98	0.21
	Lung Covid-19	X-Ray	0.28	0.27
	CRAG	Pathology	0.14	0.38
	GLAS	Pathology	0.41	0.52
	CPM-15	Pathology	0.56	0.58
	Kumar	Pathology	0.03	0.31
Dice Score Coefficient	Abdomen Covid-19	CT	0.364	0.977
	Lung Covid-19	X-Ray	0.138	0.258
	CRAG	Pathology	0.066	0.257
	GLAS	Pathology	0.098	0.452
	CPM-15	Pathology	0.476	0.63
	Kumar	Pathology	0.015	0.336

performance of SAM and Fast-SAM over these selected medical datasets. This is the issue that we look forward to working on. However, we propose using Fast-SAM's architecture and Med-SAM's weights. We analyzed the respective models' latency and precision performance for that process by taking inferences from these available datasets. Adding more datasets and testing over larger datasets can reveal more issues, as a pattern can be observed: they perform very differently under different conditions. Inference these models to just create the mask proposal is not only the key to better performance. We follow that other datasets require their way for better performance. For example, the prompts differ for every dataset and cannot be unified directly. We mostly stressed text prompts due to ease of use and practicality. Thus, similar to Fast-SAM, we show that the resulting lightweight image encoder can be automatically compatible with the mask decoder in the original SAM. And identical to MedSAM, this work attempts to adapt SAM to medical image segmentation by fine-tuning the pretrained model on medical image datasets. We look forward to working with the community to advance this exciting research area.

Software and Data

The codes (notebooks) and datasets used are publicly available at https://github.com/Pushkar1853/FaM-SAM. The notebooks consist of inference notebooks for the Fast-SAM, SAM, and Med-SAM on the selected datasets. The used metrics are also discussed, and all the results discussed above are replicable and reproducible in similar conditions.

References

[1] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," arXiv preprint arXiv:2304.02643, 2023.

[2] Zhao, X., Ding, W., An, Y., Du, Y., Yu, T., Li, M., Tang, M., Wang, J. (2023). Fast Segment Anything. ArXiv, abs/2306.12156.

[3] Zhang, Y., Jiao, R. (2023). How Segment Anything Model (SAM) Boost Medical Image Segmentation? ArXiv, abs/2305.03678.

[4] Hu and X. Li, "When sam meets medical images: An investigation of segment anything model (sam) on multi-phase liver tumor segmentation," arXiv preprint arXiv:2304.08506, 2023.

[5] B. Wang, A. Aboah, Z. Zhang, and U. Bagci, "Gazesam: What you see is what you segment," arXiv preprint arXiv:2304.13844, 2023.

[6] S. He, R. Bao, J. Li, P. E. Grant, and Y. Ou, "Accuracy of segment anything model (sam) in medical image segmentation tasks," arXiv preprint arXiv:2304.09324, 2023.

[7] A. M. Maciej, D. Haoyu, G. Hanxue, Y. Jichen, "Segment anything model for medical image analysis: an experimental study," arXiv preprint arXiv:2304.10517, 2023.

[8] D. Cheng, Z. Qin, Z. Jiang, S. Zhang, Q. Lao, and K. Li, "Sam on medical images: A comprehensive study on three prompt modes," arXiv preprint arXiv:2305.00035, 2023.

[9] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, "Segment anything model for medical images?" arXiv preprint arXiv:2304.14660, 2023.

[10] G.-P. Ji, D.-P. Fan, P. Xu, M.-M. Cheng, B. Zhou, and L. V. Gool, "Sam struggles in concealed scenes - an empirical study on "segment anything"," arXiv preprint arXiv:2304.06022, 2023.

[11] W. Ji, J. Li, Q. Bi, W. Li, and L. Cheng, "Segment anything is not always perfect: An investigation of sam on different real-world applications," arXiv preprint arXiv:2304.05750, 2023.

[12] Y. Liu, J. Zhang, Z. She, A. Kheradmand, and M. Armand, "Samm (segment any medical model): A 3d slicer integration to sam," arXiv preprint arXiv:2304.05622, 2023.

[13] M. Hu, Y. Li, and X. Yang, "Skinsam: Empowering skin cancer segmentation with segment anything model," arXiv preprint arXiv:2304.13973, 2023.

[14] Y. Li, M. Hu, and X. Yang, "Polyp-sam: Transfer sam for polyp segmentation," arXiv preprint arXiv:2305.00293, 2023.

[15] J. Ma and B. Wang, "Segment anything in medical images," arXiv preprint arXiv:2304.12306, 2023.

[16] J. Wu, R. Fu, H. Fang, Y. Liu, Z. Wang, Y. Xu, Y. Jin, and T. Arbel, "Medical sam adapter: Adapting segment anything model for medical image segmentation," arXiv preprint arXiv:2304.12620, 2023.

[17] Ke, L., Ye, M., Danelljan, M., Liu, Y. (2023). Segment Anything in High Quality. ArXiv, abs/2306.01567.

[18] Zhang, C., Han, D., Qiao, Y., Kim, J.U., Bae, S. (2023). Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. ArXiv, abs/2306.14289.

[19] Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., Chi, H., Hu, X., Fan, D., Dong, F., Ni, D. (2023). Segment Anything Model for Medical Images? ArXiv, abs/2304.14660.

Appendix

1. **Dice Score Coefficient** : the Dice Coefficient is 2 * the Area of Overlap divided by the total number of pixels in both images. Dice similarity coefficient is a spatial overlap index and a reproducibility validation metric. The value of a DSC ranges from 0, indicating no spatial overlap between two sets of binary segmentation results, to 1, indicating complete overlap.

```
def DICE_COE(mask1, mask2):
    intersect = np.sum(mask1*mask2)
    dice = (2 * intersect ) / (np.sum(mask1) + np.sum(mask2))
    dice = np.mean(dice)
    return dice
```

2. **Pixel Accuracy**: Pixel Accuracy (PA) is a semantic segmentation metric that denotes the percent of pixels that are accurately classified in the image. This metric calculates the ratio between the amount of adequately classified pixels and the total number of pixels in the image as.

```
def mean_pixel_accuracy(pixel_correct, pixel_labeled):
    numerator = 1.0 * np.sum(pixel_correct)
    denominator = (np.spacing(1) + np.sum(pixel_labeled))
    mean_pixel_accuracy = numerator/ denominator
    return mean_pixel_accuracy
```

3. **Mean Intersection of Union**: mIoU is the ratio of the intersection of the two boxes' areas to their combined areas. The ground truth bounding box and the anticipated bounding box both encompass the area of union, which is the denominator.

```
def compute_iou(mask1, mask2):
    intersection = np.logical_and(mask1, mask2)
    union = np.logical_or(mask1, mask2)
    iou = np.sum(intersection) / np.sum(union)
    iou = round(iou, 2)
    return iou
```

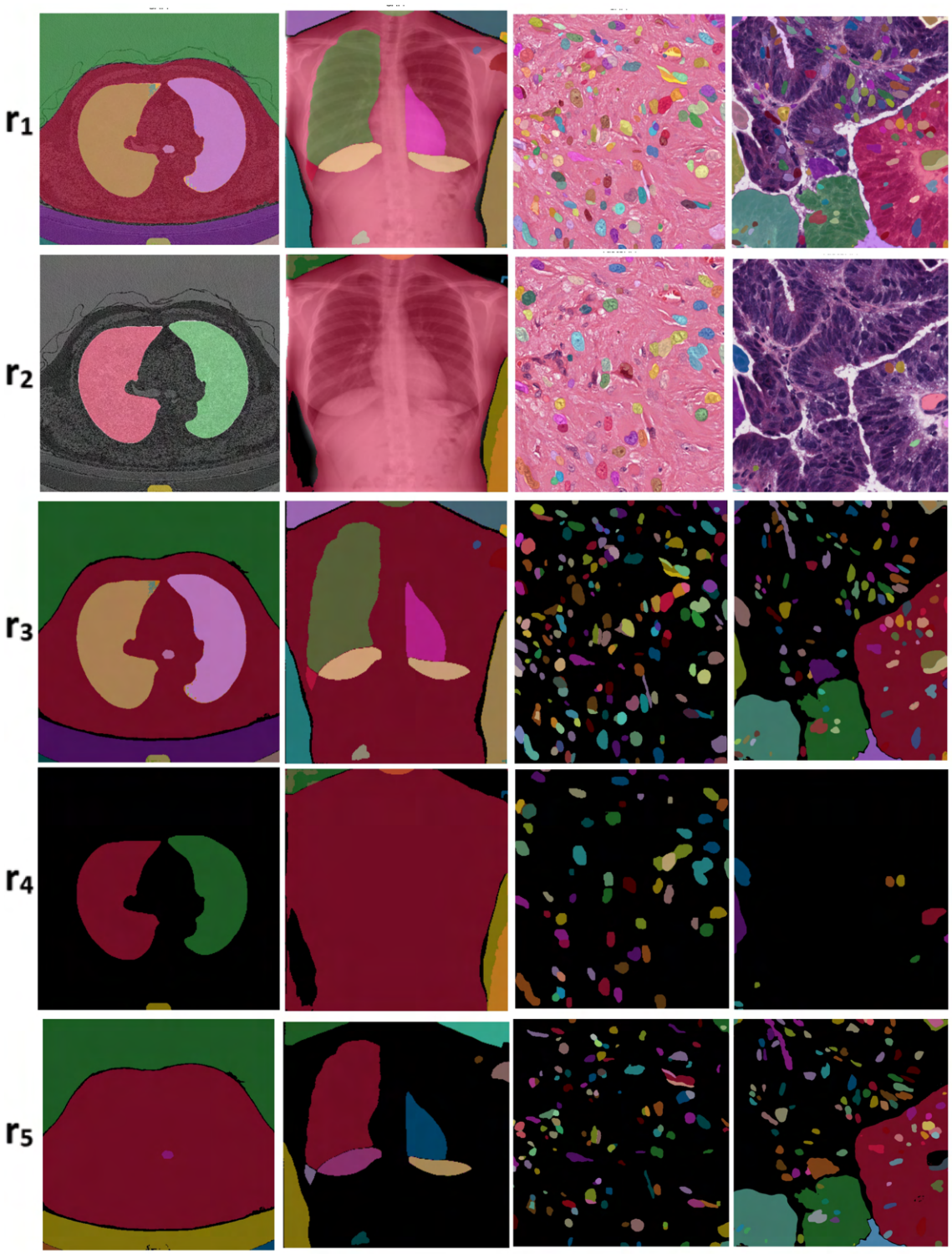

Figure 6. Example results on diverse datasets (r: row). r1, r3: Fast-SAM, r2,r4: SAM, r5: Difference in the masks. The dealt datasets (c: column). c1: Abdomen CT, c2: Lung X-Ray, c3: CPM-15, c4: Kumar Pathological Images.Obtained by segment-everything mode and giving text-prompts on necessary occasions.

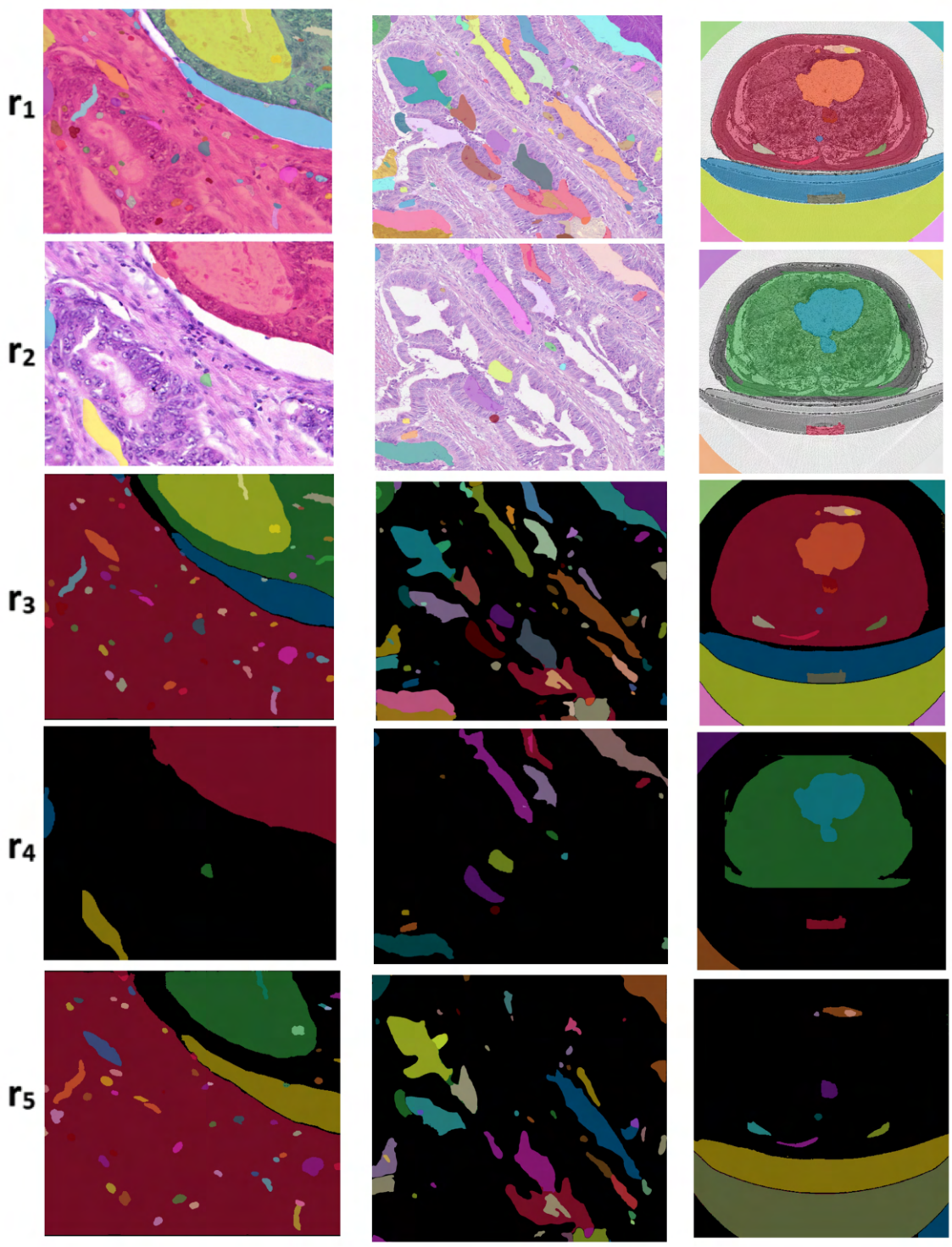

Figure 7. Example results on diverse datasets (r: row). r1, r3: Fast-SAM, r2,r4: SAM, r5: Difference in the masks. The dealt datasets (c: column). c1: GLAS Pathology, c2: CRAG Pathology, c3: Abdomen CT. Obtained by segment-everything mode and giving text-prompts on necessary occasions.