


# Supplementary - S-SAM: SVD-based Fine-Tuning of Segment Anything Model for Medical Image Segmentation

Jay N. Paranjape<sup>1</sup>, Shameema Sikder<sup>2,3</sup>, S. Swaroop Vedula<sup>3</sup>, and Vishal M. Patel<sup>1</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, USA

[jparanj1@jhu.edu](mailto:jparanj1@jhu.edu)

<sup>2</sup> Wilmer Eye Institute, The Johns Hopkins University, Baltimore, USA

<sup>3</sup> Malone Center for Engineering in Healthcare, The Johns Hopkins University, Baltimore, USA

**Preliminaries:** Segment Anything Model or SAM [3] was recently proposed as a foundational model for prompt-guided image segmentation. These prompts can be (i) foreground/background points that are vectorized using positional embeddings, (ii) bounding boxes that can be represented by the positional embeddings of their corner points, (iii) masks that can be encoded through a Convolutional Neural Network (CNN) or (iv) text that uses CLIP [4] embeddings. Note that support for the text-based prompts is not included in SAM’s released codebase. Prompt-guided learning is facilitated through separate encoders for the image and the prompt, which are then fused using a mask decoder module. The image encoder in SAM is a Vision Transformer (ViT) [1] that is pre-trained using the Masked Auto-Encoder (MAE) strategy [2] and is responsible for the majority of the memory consumption. The mask decoder is a lightweight transformer-based component while the prompt encoder embeds the different prompts as described earlier and combines them.

**Number of Singular Values Tuned:** We conduct experiments to check whether all the singular values need to be tuned to model the required domain shift. As seen in Table 1, we tune only the top  $k\%$  values, where  $k \in \{1, 10, 50, 100\}$ . However, we see a significant drop in performance on reducing  $k$  on the CholecSeg8k dataset. This is expected since the eigenvectors corresponding to the top singular values for the natural image domain might not be the most relevant vectors for the new medical domain. Hence, best performance is observed when  $k$  is 100.

**Experimental Setup:** We use the ‘ViT-base’ backbone checkpoint from SAM for initializing our model, while the weights of the TAL network are initialized using the default settings of Pytorch (Kaiming Uniform). We apply augmentations including random rotation ( $\pm 10^\circ$ ) with 0.5 probability, random saturation change with a scale of 2 with 0.2 probability, and random brightness change with a scale of 2 with 0.5 probability during training for all input images, followed by

Table 1: Ablation analysis on percentage of singular values tuned.

| Percent of singular values tuned | Avg. DSC |
|----------------------------------|----------|
| 1                                | 0.49     |
| 10                               | 0.54     |
| 50                               | 0.56     |
| 100                              | 0.71     |

the normalization used in SAM. All training is done with the AdamW optimizer with a learning rate of 1e-4, on a single Nvidia RTX A6000 GPU. The memory requirement for a given training instance is less than 12 GB when the image is resized to  $256 \times 256$ . The loss function used for all experiments is the sum of dice loss and focal loss between the ground truth label and the predicted mask.

Table 2: Results on Abdominal Ultrasound.

| Method                        | Objectwise DSC |        |          |         |          |              |       |        |             |
|-------------------------------|----------------|--------|----------|---------|----------|--------------|-------|--------|-------------|
|                               | Liver          | Kidney | Pancreas | Vessels | Adrenals | Gall Bladder | Bones | Spleen | Avg.        |
| <b>Traditional DL methods</b> |                |        |          |         |          |              |       |        |             |
| UNet                          | 0.28           | 0.37   | 0.11     | 0.16    | 0.85     | 0.08         | 0.17  | 0.14   | 0.27        |
| TransUNet                     | 0.18           | 0.09   | 0.03     | 0.03    | 0        | 0.11         | 0.05  | 0.02   | 0.08        |
| MedT [5]                      | 0.18           | 0.03   | 0.27     | 0.10    | 0.85     | 0.15         | 0.02  | 0.08   | 0.21        |
| <b>SAM based methods</b>      |                |        |          |         |          |              |       |        |             |
| SAM w/ text prompt            | 0.17           | 0.20   | 0.72     | 0.21    | 0.44     | 0.65         | 0.67  | 0.63   | 0.46        |
| SAM w/ point prompt           | 0.11           | 0      | 0.01     | 0       | 0.01     | 0.01         | 0.01  | 0.01   | 0.02        |
| SAM with full finetuning      | 0.21           | 0.48   | 0.67     | 0.56    | 0.81     | 0.69         | 0.54  | 0.53   | 0.56        |
| MedSAM                        | 0.14           | 0.03   | 0.01     | 0.01    | 0        | 0.01         | 0     | 0.02   | 0.03        |
| SAMed                         | 0.20           | 0.50   | 0.61     | 0.56    | 0.82     | 0.63         | 0.54  | 0.54   | 0.55        |
| AdaptiveSAM                   | 0.36           | 0.30   | 0.50     | 0.40    | 0.86     | 0.63         | 0.67  | 0.54   | 0.53        |
| Low Rank Adaptation of SAM    | 0.43           | 0.35   | 0.45     | 0.61    | 0.90     | 0.59         | 0.67  | 0.67   | 0.58        |
| S-SAM (Ours)                  | 0.32           | 0.52   | 0.80     | 0.61    | 0.91     | 0.75         | 0.67  | 0.43   | <b>0.63</b> |

Table 3: Results on ChestXDet.

| Method                        | Object wise DSC |      |      |       |      |      |      |      |      |      |      |      |      |             |
|-------------------------------|-----------------|------|------|-------|------|------|------|------|------|------|------|------|------|-------------|
|                               | Ef              | No   | Cm   | Fb    | Co   | Em   | Ma   | Ca   | Pt   | Pn   | Fr   | At   | Dn   | Avg.        |
| <b>Traditional DL methods</b> |                 |      |      |       |      |      |      |      |      |      |      |      |      |             |
| UNet                          | 0.15            | 0.08 | 0.06 | 0     | 0.13 | 0.02 | 0.95 | 0    | 0.08 | 0    | 0.50 | 0.02 | 0.02 | 0.15        |
| TransUNet                     | 0.06            | 0.87 | 0.06 | 0.59  | 0.13 | 0.01 | 0.89 | 0    | 0.74 | 0    | 0.08 | 0    | 0    | 0.26        |
| MedT                          | 0.06            | 0.75 | 0.08 | 0.01  | 0.10 | 0.03 | 0.12 | 0    | 0.91 | 0    | 0    | 0.37 | 0.07 | 0.19        |
| <b>SAM based methods</b>      |                 |      |      |       |      |      |      |      |      |      |      |      |      |             |
| SAM w/ text prompt            | 0.05            | 0.13 | 0.53 | 0.36  | 0.15 | 0.28 | 0.23 | 0.10 | 0.37 | 0.07 | 0.40 | 0    | 0.26 | 0.22        |
| SAM w/ point prompt           | 0.04            | 0    | 0.01 | 0.01  | 0.04 | 0.01 | 0    | 0    | 0    | 0    | 0    | 0    | 0    | 0.02        |
| SAM with full finetuning      | 0.55            | 0.88 | 0.87 | 0.086 | 0.52 | 0.93 | 0.95 | 0.93 | 0.84 | 0.93 | 0.86 | 0.92 | 0.94 | <b>0.84</b> |
| MedSAM                        | 0.04            | 0    | 0.02 | 0.01  | 0.04 | 0.02 | 0    | 0    | 0    | 0    | 0.02 | 0    | 0.02 | 0.01        |
| SAMed                         | 0.50            | 0.89 | 0.90 | 0.83  | 0.50 | 0.93 | 0.94 | 0.91 | 0.83 | 0.92 | 0.85 | 0.93 | 0.93 | 0.83        |
| AdaptiveSAM                   | 0.52            | 0.88 | 0.86 | 0.86  | 0.43 | 0.93 | 0.95 | 0.91 | 0.84 | 0.93 | 0.86 | 0.94 | 0.93 | 0.83        |
| Low Rank Adaptation of SAM    | 0.50            | 0.89 | 0.90 | 0.83  | 0.42 | 0.93 | 0.96 | 0.91 | 0.84 | 0.93 | 0.86 | 0.94 | 0.94 | 0.83        |
| S-SAM (Ours)                  | 0.48            | 0.89 | 0.87 | 0.86  | 0.4  | 0.93 | 0.96 | 0.94 | 0.84 | 0.94 | 0.87 | 0.94 | 0.95 | <b>0.84</b> |

## References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Housby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
2. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021)
3. 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)
4. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021)
5. Valanarasu, J.M.J., Oza, P., Hacihaliloglu, I., Patel, V.M.: Medical transformer: Gated axial-attention for medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. pp. 36–46. Springer International Publishing, Cham (2021)