Supplementary Material For WACV2024
SAM Fewshot Finetuning for Anatomical Segmentation in Medical Images

Weiyi Xie, Nathalie Willems, Shubham Patil, Yang Li, and Mayank Kumar
Stryker AI Research
{weiyi.xie,nathalie.willems,shubham.patil,yang.li1,mayank.kumar}@stryker.com

1. Embedding Analysis

We present t-SNE plots as a means to visualize feature embeddings generated by the image encoder in SAM across all datasets utilized in this study. These t-SNE plots were created using a perplexity of 25 and 5000 iterations. We sampled equal number (n=50) of feature embeddings per class for each task on the test set images from SAM's image encoder. As illustrated in Figure 1, the majority of classes exhibit clear separation in the embedding space. However, for the knee classes (tibia and femur), the separation within the embedding space is comparatively less distinct. This observation suggests the presence of ambiguities when utilizing SAM's feature embedding to distinguish between femur and tibia. Nonetheless, it is worth noting that the proposed mask decoder and fine-tuning strategy appear to mitigate these ambiguities effectively. This is evidenced by the relatively high segmentation performance achieved for tibia and femur after fine-tuning, especially when compared to the results obtained through the fully-supervised method (refer to Table 2).

2. Number of Labeled Images in Fine-Tuning

In this section, we undertake a comparative analysis of the few-shot fine-tuning method, leveraging 5, 20, 50, 100, and 200 labeled images, in order to extend our investigation, as initially presented in Table 2. The results, depicted in Figure 2, clearly demonstrate that the inclusion of a greater number of labeled images generally enhances segmentation performance, as assessed through the IoU and ASSD metrics. Notably, when considering femur and tibia, the IoU achieved with 200 labeled images stands at an impressive 98.1 % and 97.9 %, surpassing the performance of the fully-supervised method, which records 97.8 % and 97.5 %, respectively. For the segmentation of the left atrium, employing 200 labeled images yields an IoU of 89.6 %, compared to the fully-supervised method's IoU of 88.9 %. Similarly, when focusing on vertebrae segmentation, employing 200 labeled images results in an IoU of 93.1%, surpassing the fully-supervised method's performance at 92.7%. It is worth noting, however, that this trend does not hold uniformly for all anatomical structures investigated in this study. For postcava and aorta, achieving fully-supervised performance may require adding more labeled images than 200 in the fine-tuning process.

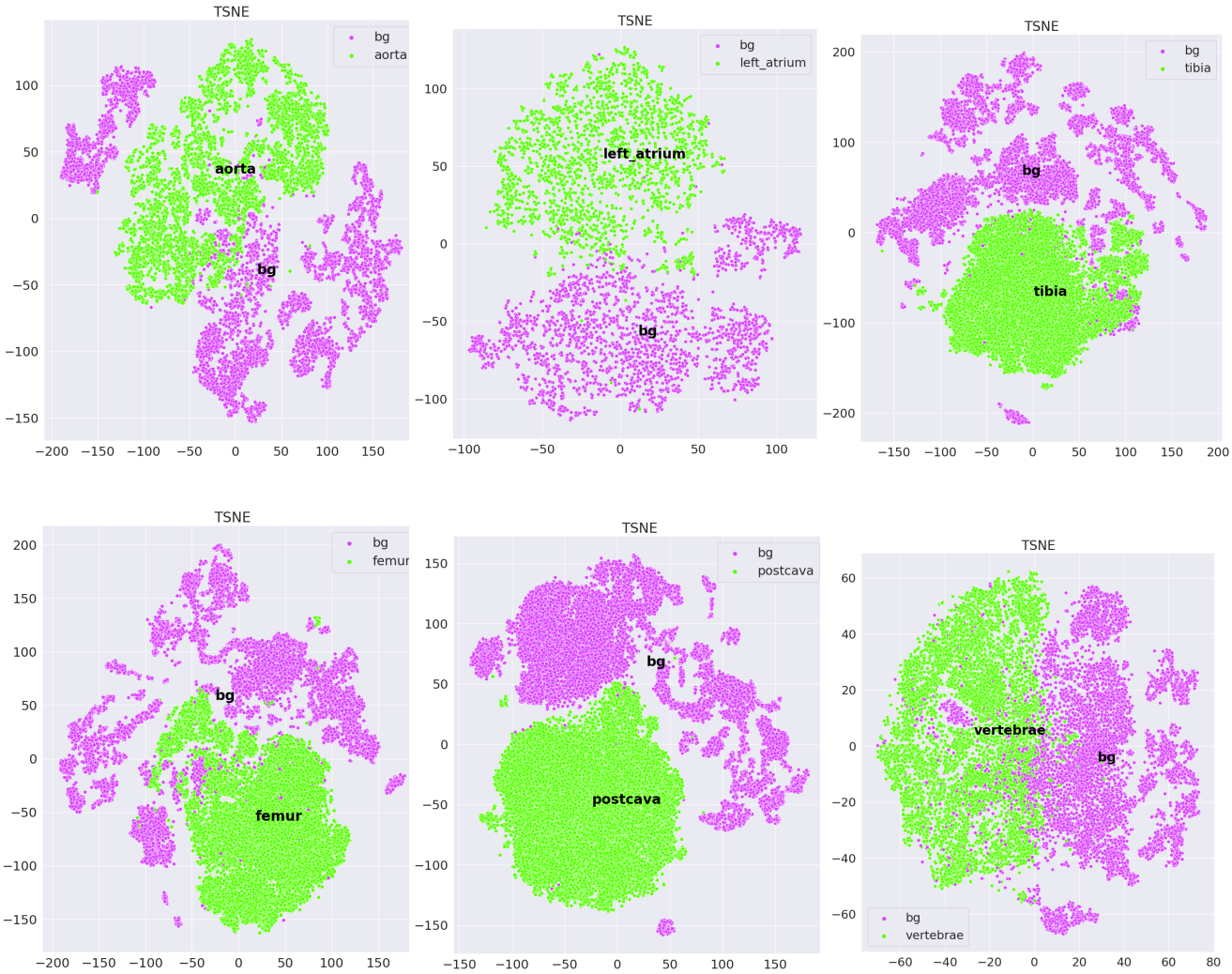

Figure 1. t-SNE plots showing the feature embeddings generated by the image encoder in the SAM for separating the anatomy against the background among tasks. The first row shows the class separation for aorta, left atrium and tibia from the left to the right, while the second row shows femur, postcava, and vertebrae class separation from the left to the right.

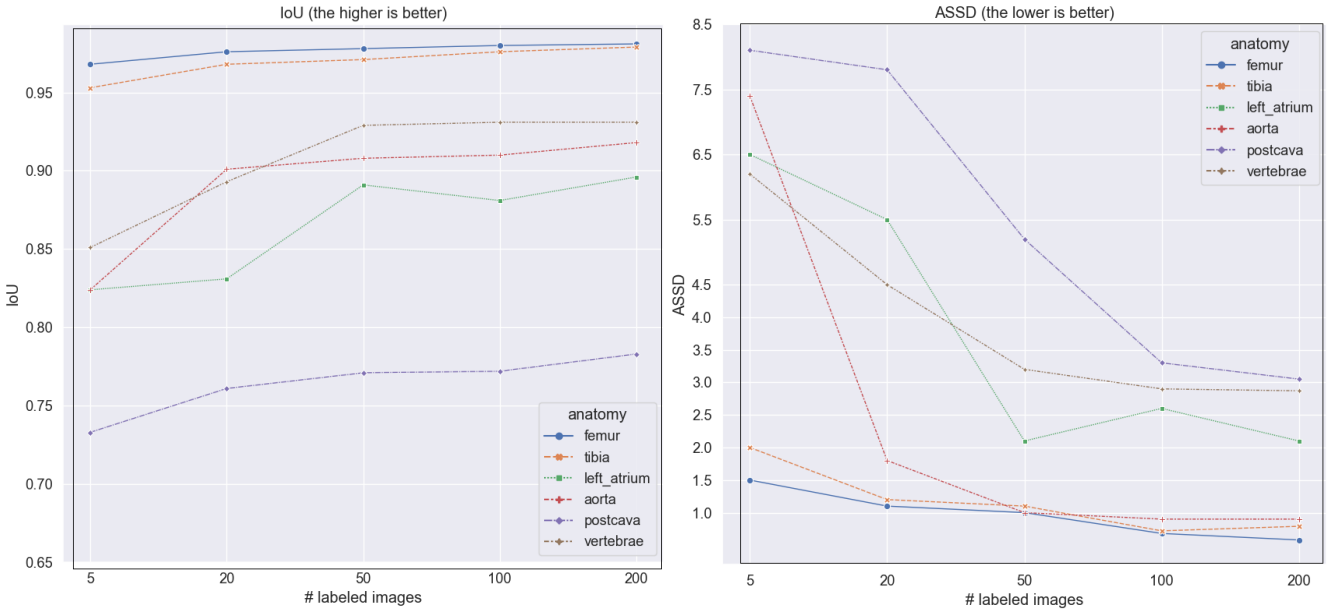

Figure 2. Incorporating a greater number of labeled images during the fine-tuning process generally enhances segmentation performance across various anatomical structures in this study. This improvement is manifested by an increase in Intersection over Union (IoU) on the left and a reduction in Average Surface Symmetric Distance (ASSD) on the right.