MedLSAM: Localize and Segment Anything Model for 3D Medical Images

Wenhui Lei[1,2], Wei Xu[3,4], Xiaofan Zhang[1,2*], Kang Li[4], and Shaoting Zhang[1]

[1]Shanghai AI Lab, Shanghai, China
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]School of Biomedical Engineering & Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China
[4]West China Biomedical Big Data Center, West China Hospital, Sichuan University, Chengdu, China
xiaofan.zhang@sjtu.edu.cn

Abstract. The Segment Anything Model (SAM) has recently emerged as a groundbreaking model in the field of image segmentation. Nevertheless, both the original SAM and its medical adaptations necessitate slice-by-slice annotations, which directly increase the annotation workload with the size of the dataset. We propose MedLSAM to address this issue, ensuring a constant annotation workload irrespective of dataset size and thereby simplifying the annotation process. Our model introduces a few-shot localization framework capable of localizing any target anatomical part within the body. To achieve this, we develop a Localize Anything Model for 3D Medical Images (MedLAM), utilizing two self-supervision tasks: relative distance regression (RDR) and multi-scale similarity (MSS) across a comprehensive dataset of 14,012 CT scans. We then establish a methodology for accurate segmentation by integrating MedLAM with SAM. By annotating only six extreme points across three directions on a few templates, our model can autonomously identify the target anatomical region on all data scheduled for annotation. This allows our framework to generate a 2D bounding box for every slice of the image, which are then leveraged by SAM to carry out segmentations. We conducted experiments on two 3D datasets covering 38 organs and found that MedLSAM matches the performance of SAM and its medical adaptations while requiring only minimal extreme point annotations for the entire dataset. Furthermore, MedLAM has the potential to be seamlessly integrated with future 3D SAM models, paving the way for enhanced performance. Our code is public at https://github.com/openmedlab/MedLSAM.

Keywords: SAM · Medical Image Segmentation · Contrastive Learning

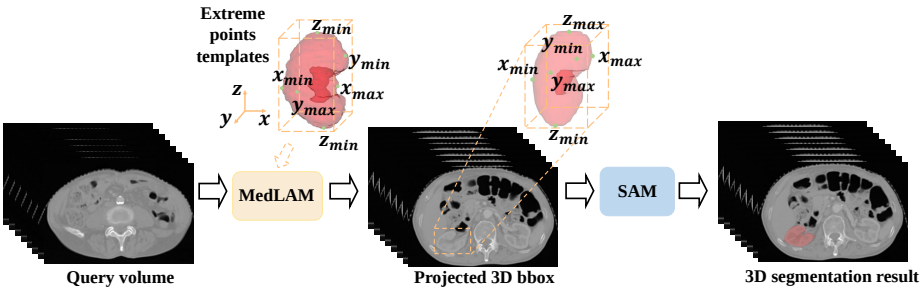

Fig. 1: The overall segmentation pipeline of MedLSAM operates as follows. Given a dataset of any size, MedLSAM first applies a localization process (MedLAM) to identify the six extreme points (in the z, x, and y directions) of any anatomical region of interest. This process results in the generation of a 3D bounding box encompassing the targeted organ or structure. Subsequently, for each slice within this 3D bounding box, a corresponding 2D bounding box is generated. These 2D bounding boxes are then utilized by the Segment Anything Model (SAM) to carry out precise segmentation of the target anatomy, thereby automating the entire segmentation process.

1 Introduction

The Segment Anything Model (SAM) [14] has recently demonstrated remarkable capabilities in a broad range of segmentation tasks due to its ability to manage diverse objects. Previous research has explored the application of SAM to medical image segmentation [5, 8, 9, 11, 20, 22, 25, 28, 29]. Particularly with fine-tuned medical adaptations [20, 25], SAM models have achieved impressive performance in this specialized area. However, SAM's application to medical image segmentation poses a significant challenge due to its requirement for manual annotations. These include labeled points or bounding boxes that delineate the segmentation region, which are both time-intensive and costly to produce.

In this paper, we introduce MedLSAM, an automated medical image segmentation model designed to significantly reduce the annotation workload. As illustrated in Fig. 1, MedLSAM employs a two-stage methodology. The first stage involves a few-shot localization framework that automatically identifies the positions of target organs within volumetric medical images. The subsequent stage utilizes the bounding boxes generated in the first stage by applying the SAM model to execute precise image segmentation. The result is a fully autonomous pipeline that eliminates the need for manual intervention.

The few-shot localization framework, MedLAM, is an extension of our previous work [17] and is premised on the observation that the spatial distribution of organs maintains strong similarities across different individuals. Our previous studies trained the model on a relatively smaller dataset. In contrast, this current study significantly expands the dataset to include 14,012 CT scans from 16

* Corresponding author.

different datasets. This allows us to train a unified, comprehensive model capable of localizing structures across the entire body. The training process involves a projection network that predicts the 3D physical offsets between any two patches within the same image, thereby mapping every part of the scans onto a shared 3D latent coordinate system.

While our localization approach is robust, we acknowledge that individual variations in anatomical positioning could result in different anatomical structures sharing the same latent coordinates across various images. To mitigate this issue, we refine our localization accuracy by extracting pixel-level features from our points of interest. This method enables us to identify the most similar feature within the vicinity of the initially localized point, thus enhancing the overall localization accuracy. Our approach is inspired by the self-supervised learning tasks proposed in [26, 27], which strive to maximize the similarity between the original and augmented instances of the same image point and minimize the similarity between different points. This technique improves the accuracy and precision of our model's localization.

For the segmentation stage, we employ the original SAM and the well-established MedSAM [20] as the foundation for our segmentation process. MedSAM, previously fine-tuned on a comprehensive collection of medical image datasets, has exhibited considerable performance in 2D and 3D medical image segmentation tasks. The use of such a robust model bolsters the reliability and effectiveness of our proposed pipeline.

The effectiveness of MedLSAM is validated through experiments on two 3D datasets comprising 38 organs. The results demonstrate that MedLSAM parallels the performance of SAM and its medical adaptations while significantly reducing the burden of manual annotation.

2 Methodology

In this section, we elucidate the mechanisms underpinning MedLAM's functionality. Initially, in Section 2.1, we explore the training of MedLAM, during which global anatomical coordinates and local image features are extracted from any given point within a scan, aiding in identifying the most similar point within a query scan. Subsequently, in Section 2.2, we detail the inference process of MedLAM and its integration with the Segment Anything Model (SAM) to give rise to the comprehensive MedLSAM model.

2.1 Training of MedLAM

Our MedLAM model, as illustrated in Fig. 2, comprises two main components: Relative Distance Regression (RDR) and Multi Scale Similarity (MSS).

We start by selecting a volumetric image $\boldsymbol{v}$ from the unannotated training set. We then extract two large image patches from $\boldsymbol{v}$, which serve as the source for further extractions. These large patches undergo a variety of transformations to produce two pairs of patches, namely, the original patch pair $(\boldsymbol{x_q}, \boldsymbol{x_s})$ and the transformed patch pair $(\boldsymbol{x'_q}, \boldsymbol{x'_s})$.

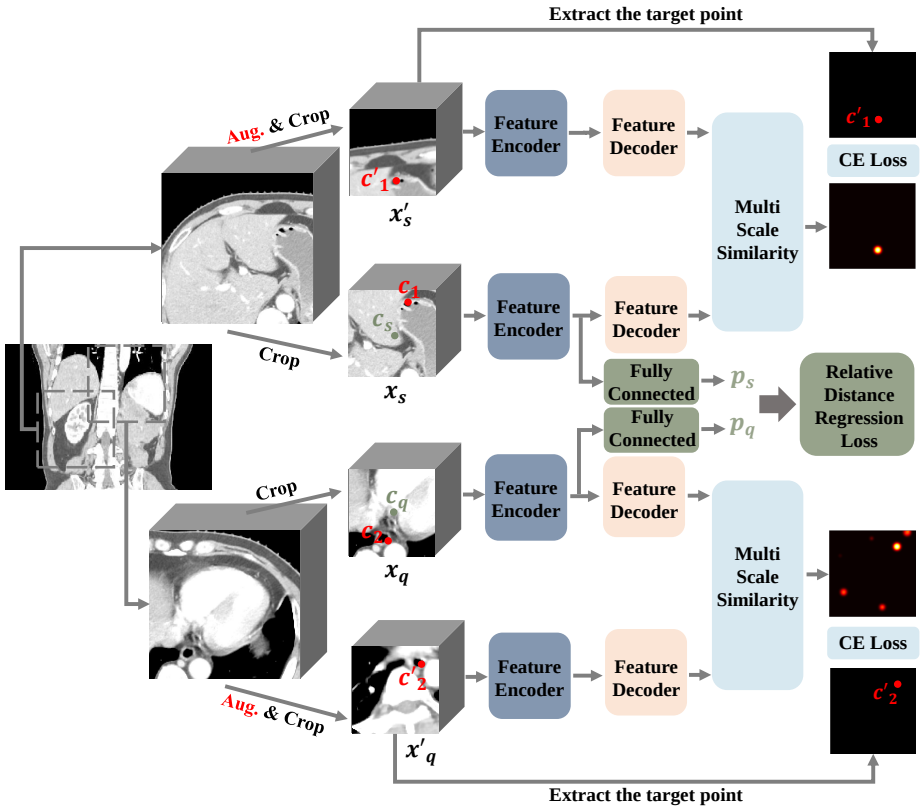

Fig. 2: The learning process of MedLAM.

Relative Distance Regression (RDR) In this step, we leverage the RDR methodology, an extension of our previous work [17]. This involves mapping 3D scan images from different individuals onto a unified implicit 3D anatomical coordinate system, ensuring that identical anatomical structures from different individuals share the same coordinate. As a result, it allows us to perform an initial, coarse localization of the point within a query scan that shares the same implicit coordinate as our point of interest.

The RDR model aims to predict the 3D offset between the query patch $\boldsymbol{x_q}$ and the support patch $\boldsymbol{x_s}$. Considering $\boldsymbol{e} \in R^3$ as the pixel spacing of $\boldsymbol{v}$, and $\boldsymbol{c_q}, \boldsymbol{c_s} \in R^3$ as the centroid coordinates of $\boldsymbol{x_q}$ and $\boldsymbol{x_s}$ in $\boldsymbol{v}$ respectively, the ground truth offset $\boldsymbol{d'_{qs}}$ from $\boldsymbol{x_q}$ to $\boldsymbol{x_s}$ in the physical space can be calculated as:

$$\boldsymbol{d'_{qs}} = (\boldsymbol{c_s} - \boldsymbol{c_q}) \cdot \boldsymbol{e} \tag{1}$$

Both $\boldsymbol{x_s}$ and $\boldsymbol{x_q}$ undergo processing via an encoder to distill high-level features. Subsequently, fully connected layers map these features to their corresponding 3D latent vectors, $\boldsymbol{p_s}$ and $\boldsymbol{p_q}$, each $\in R^3$. This process leads to the

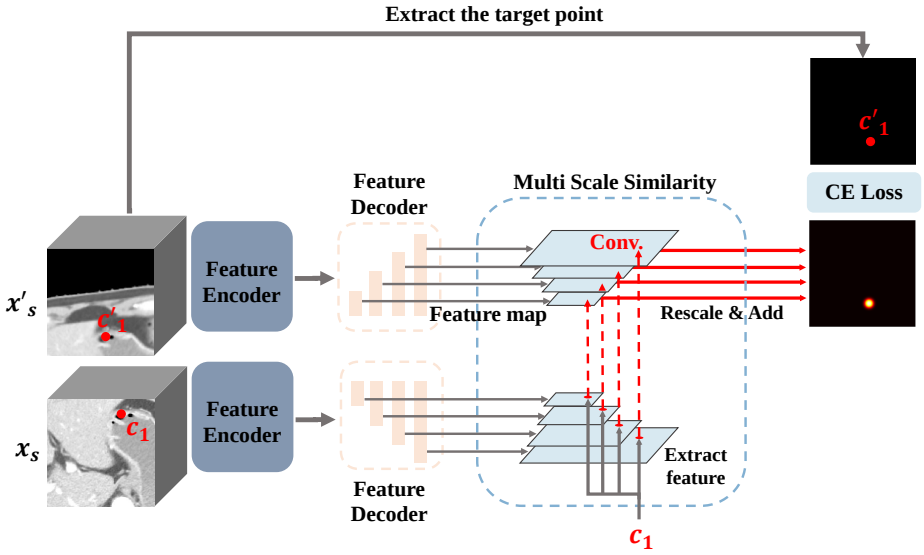

Fig. 3: Deatils of the Multi Scale Similarity (MSS).

predicted offset $\boldsymbol{d_{qs}} \in R^3$ from the query patch $\boldsymbol{x_q}$ to the support patch $\boldsymbol{x_s}$ being computed as:

$$\boldsymbol{d_{qs}} = r \cdot tanh(\boldsymbol{p_s} - \boldsymbol{p_q}) \tag{2}$$

The utilization of the hyperbolic tangent function $tanh$ in conjunction with the hyper-parameter r is intended to dictate the upper and lower bound of $\boldsymbol{d_{qs}}$, thereby covering the largest feasible offset. Lastly, to measure the difference between $\boldsymbol{d_{qs}}$ and $\boldsymbol{d'_{qs}}$, we employ the Mean Square Error (MSE) loss function:

$$L_D = ||\boldsymbol{d_{qs}} - \boldsymbol{d'_{qs}}||^2 \tag{3}$$

Multi Scale Similarity (MSS) Given the inherent variations in anatomical positioning across different individuals, regions sharing the same latent coordinates in various images may still correspond to different anatomical structures. Therefore, we need to further refine the precision of our localization by extracting local pixel-level features from our points of interest. This allows us to pinpoint the most similar feature within the vicinity of the initially localized point, thereby enhancing the overall localization accuracy. This is inspired by the work in [26, 27], which ensures that augmented instances of the same image yield highly similar features for the same point, while different points exhibit substantially divergent features.

More specifically, as shown in Fig. 3, the inputs to our MSS process include multi-scale feature maps extracted from $\boldsymbol{x_s}$ and $\boldsymbol{x'_s}$, along with a chosen point

c_1 from $\boldsymbol{x_s}$, whose corresponding point in $\boldsymbol{x'_s}$ is the c'_1. We extract the feature vectors corresponding to point c_1 from the various scale feature maps of $\boldsymbol{x_s}$, and we compute the similarity between these feature vectors and the corresponding scale feature maps in $\boldsymbol{x'_s}$. After resizing the resulting similarity maps to the original image size, we aggregate them. This process allows us to pinpoint the location within $\boldsymbol{x'_s}$ that exhibits the highest similarity to point c_1, thereby further refining our localization.

2.2 Inference of MedLSAM

The inference stage for our MedLSAM framework combines the strengths of MedLAM for landmark localization and MedSAM for medical image segmentation, as shown in Fig. 4.

Initially, we utilize MedLAM to localize the desired landmark in the query image. We conceptualize the localization task as maneuvering an agent from a randomly initialized position towards the target location. A patch is extracted from a random position within the query image, and simultaneously, a support patch is extracted from the support image, centered around the pre-specified landmark. Upon processing these two patches through the MedLAM model, we obtain a 3D offset that represents the estimated relative spatial displacement between the query and target positions. By updating the agent's location based on this offset, we achieve a coarse localization of the landmark within the query image.

For refining the landmark localization, the Multi Scale Similarity (MSS) component of MedLAM is utilized. We extract multi-scale feature maps around the coarsely localized point in the query image and its corresponding point in the support image, perform similarity calculations, and aggregate the similarity maps to pinpoint the location with the highest feature similarity in the query image. This procedure significantly enhances the precision of our landmark localization.

After successfully identifying the landmarks, we transition to the segmentation stage. For this, we utilize both SAM and MedSAM, a specialized variant of SAM that has been fine-tuned for medical image datasets. Both models serve as the foundation for our segmentation tasks. The versatility of SAM and the domain-specific adaptations of MedSAM help us provide robust segmentation results, thereby adding to the overall efficacy of the MedSLAM system.

3 Experiments

3.1 Dataset

Our MedLAM model is trained on an extensive set of 16 datasets, which collectively comprise a total of 14,012 CT scans. These scans encompass various regions of the human body, providing comprehensive anatomical coverage. An

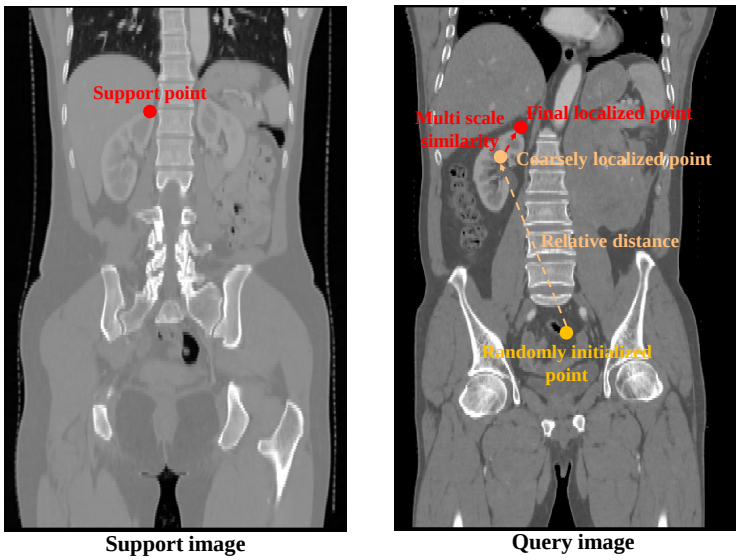

Fig. 4: Structure of our Localization Anything Model (MedLAM). $\boldsymbol{x_s}$ and $\boldsymbol{x_q}$ are the support and query patches centered at $\boldsymbol{c_s}$ and $\boldsymbol{c_q}$. We use a shared Pnet to transform $\boldsymbol{x_s}$ and $\boldsymbol{x_q}$ to 3D latent vectors $\boldsymbol{p_s}$ and $\boldsymbol{p_q}$, respectively. The Pnet contains convolution blocks to extract features and fully connected layers for projection. We apply scale factor r and hyperbolic tangent function $tanh$ to get the predicted offset $\boldsymbol{d_{qs}}$, i.e., relative position from $\boldsymbol{x_s}$ to $\boldsymbol{x_q}$.

overview of the training datasets is provided in Table 1. The diverse and abundant training data ensures the robustness and generalizability of our model across different medical imaging contexts.

To validate the effectiveness of our approach, we integrate MedLAM with two segmentation backbones: SAM [14] and MedSAM [20]. We test these combined models on two CT segmentation datasets: 1) StructSeg19 Task1 dataset for the 22 head-and-neck (HaN) organs with 50 scans; 2) the WORD dataset [19] for the 16 abdomen organs with 120 scans. For both datasets, we randomly select five scans as support volumes. For each organ in these scans, we compute the extreme coordinates and average the coordinates and features across the five images. This process generates an average representation of latent coordinates and features for each extreme point of the organ, which are then utilized in the succeeding stages of MedLAM as depicted in Sec. 2.2.

https://structseg2019.grand-challenge.org/Home/

Table 1: Detailed information of the 16 CT datasets for MedLAM training.

Dataset	Number	Anatomical Region
GLIA [3]	1338	HaN
ACRIN 6685 [18]	260	HaN
OPC-Radiomics [15]	606	HaN
Head-Neck-PET-CT [24]	298	HaN
HNSCC [7]	591	HaN/Thorax/Abdomen
autoPET [6]	1014	Whole
MELA [4]	770	Thorax
LIDC-IDRI [2]	1308	Thorax
STOIC2021 [23]	2000	Thorax
MSD-Lung [1]	95	Thorax
CBIS-DDSM [16]	2620	Thorax
AMOS 2022 [12]	500	Thorax/Abdomen
Kits19 [10]	141	Abdomen
MSD-Colon [1]	190	Abdomen
MSD-Pancreas [1]	281	Abdomen
FLARE2022 [21]	2000	Abdomen
Total	14,012	Whole

3.2 Implementation Details

Our model was trained using four NVIDIA GTX 3090 Ti GPUs. We utilized the Adam optimizer [13] with a batch size of 8, an initial learning rate of 10^{-3}, and training duration of 250 epochs.

In terms of pre-processing for MedLAM's training and testing, we rescaled the voxel spacing to [3,3,3] mm and standardized the cropping patch sizes to 64×64×64 pixels. To ensure that the scanning range was fully covered, we set the parameter r as [1500, 600, 600].

Upon utilizing SAM and MedSAM, the original images are subjected to a separate preprocessing routine in line with the standard procedures described in the original MedSAM methodology. This includes adjusting the slice resolution to $3 \times 1024 \times 1024$, normalizing the intensity.

Further, specific handling measures are adopted for different datasets based on their unique characteristics. For abdominal organs in the WORD dataset, in accordance with MedSAM, we exclude segmentation targets that consist of fewer than 100 pixels. While for the HaN organs in the StructSeg dataset, which are typically smaller, we adapt the criteria and exclude only those slices that contain fewer than 10 pixels. This adjustment ensures the model's robust performance in identifying and analyzing small yet potentially significant anatomical structures in HaN CT scans.

For the 3D bounding box obtained from MedLAM localization, we extended it by [2, 10, 10] pixels in the z, x, and y directions, respectively. This strategy ensured that the targeted organ was completely encapsulated within the box, enabling effective segmentation.

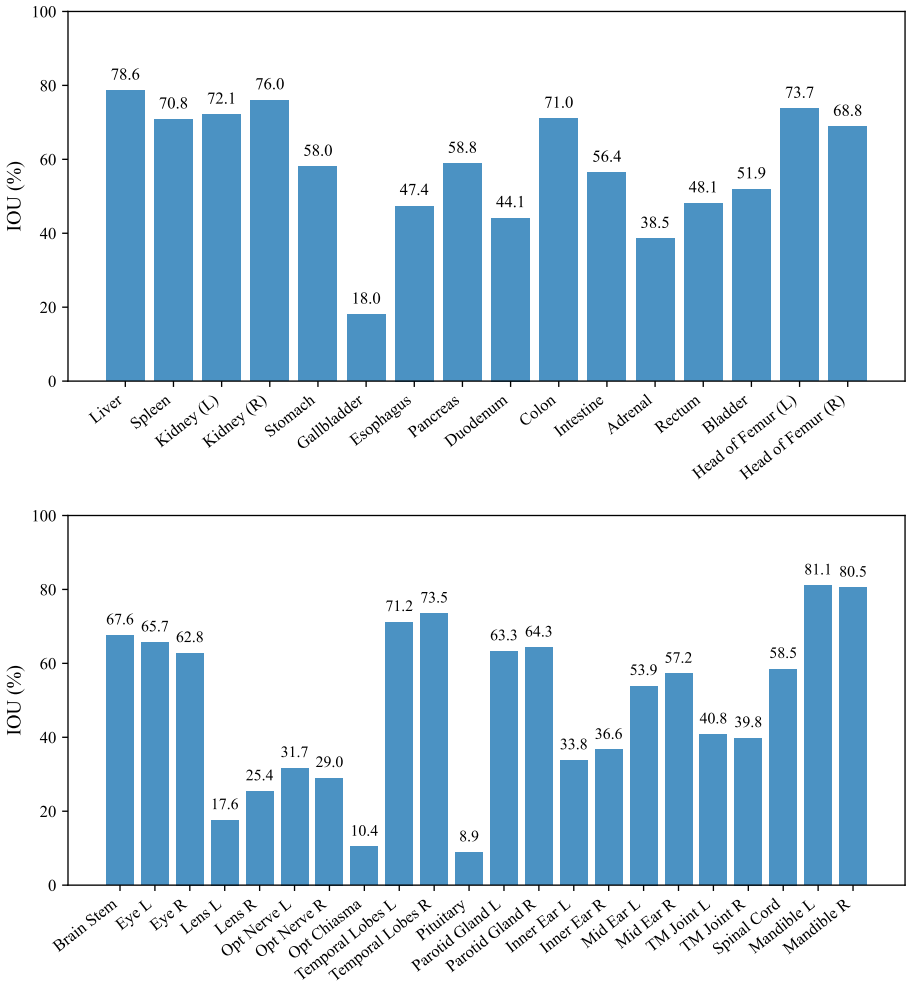

Fig. 5: The mean IOU score of each organ in the WORD (top) and StructSeg (bottom) dataset.

3.3 Experiments Results

Evaluation of LAM In this section, we begin by evaluating the localization performance of MedLAM, with the purpose of validating the viability of our universal localization model. We used the Intersection Over Union (IOU) as the metric to evaluate the accuracy of organ localization. The mean IOU of each organ is displayed in Fig. 5. In the WORD dataset, the model achieved excellent localization with the highest IOU for Head of Femur (L), reaching 0.737. Meanwhile, Gallbladder showed the lowest IOU with 0.180, suggesting room for improvement in localizing smaller organs.

Table 2: DSC (mean±std %) evaluation of 3D head-and-neck organs segmentation in the StructSeg Task1 dataset. The table compares the performance of SAM and MedSAM as a segmentation basis within the MedLsAM framework, along with results from manually assisted localizations.

Localization	MedLAM		Manual	
Organs	SAM	MedSAM	SAM	MedSAM
Brain Stem	53.5 ± 5.5	**64.7 ± 6.3**	65.2 ± 3.7	**72.8 ± 3.3**
Eye L	**63.9 ± 6.1**	61.1 ± 6.1	**67.6 ± 5.0**	66.8 ± 5.6
Eye R	**66.3 ± 5.3**	63.4 ± 5.2	**69.5 ± 4.6**	67.6 ± 4.9
Lens L	22.2 ± 7.5	**16.5 ± 3.1**	**21.4 ± 9.5**	15.9 ± 2.8
Lens R	20.6 ± 6.9	**13.6 ± 2.8**	**20.7 ± 10.8**	13.8 ± 3.5
Opt Nerve L	**31.4 ± 9.5**	29.7 ± 13.2	**32.4 ± 12.9**	22.2 ± 17.9
Opt Nerve R	**34.6 ± 8.9**	32.1 ± 12.2	32.2 ± 15.9	**36.4 ± 13.1**
Opt Chiasma	**29.0 ± 10.0**	28.9 ± 16.0	**37.9 ± 14.9**	25.3 ± 14.7
Temporal Lobes L	25.4 ± 16.3	**72.3 ± 4.8**	37.7 ± 20.2	**78.2 ± 6.4**
Temporal Lobes R	19.9 ± 20.2	**67.7 ± 8.6**	34.4 ± 21.2	**76.5 ± 7.6**
Pituitary	**36.2 ± 21.1**	28.5 ± 16.1	**36.6 ± 17.0**	29.1 ± 15.4
Parotid Gland L	7.1 ± 6.5	**44.2 ± 10.3**	27.8 ± 10.0	**50.7 ± 11.0**
Parotid Gland R	8.1 ± 8.6	**44.7 ± 8.5**	30.2 ± 9.9	**49.4 ± 10.5**
Inner Ear L	**51.7 ± 16.5**	48.1 ± 13.7	**56.4 ± 16.1**	54.7 ± 12.4
Inner Ear R	**63.8 ± 10.3**	43.5 ± 18.8	**60.5 ± 15.8**	44.5 ± 16.9
Mid Ear L	**64.1 ± 11.8**	27.7 ± 14.1	**74.4 ± 7.7**	39.5 ± 14.4
Mid Ear R	**64.7 ± 10.6**	33.2 ± 14.2	**74.4 ± 7.7**	45.0 ± 9.6
TM Joint L	**54.0 ± 8.3**	34.9 ± 12.4	**62.8 ± 11.5**	37.8 ± 16.1
TM Joint R	**58.5 ± 7.9**	43.6 ± 11.1	**64.5 ± 14.8**	46.6 ± 11.9
Spinal Cord	**9.5 ± 3.8**	9.4 ± 3.5	**40.4 ± 6.3**	32.7 ± 6.4
Mandible L	**48.3 ± 5.1**	9.0 ± 3.8	**85.3 ± 2.4**	11.1 ± 4.6
Mandible R	**43.5 ± 5.3**	2.6 ± 2.9	**80.4 ± 2.5**	12.9 ± 7.8
Average	**39.6 ± 7.6**	37.5 ± 7.1	**50.6 ± 6.3**	42.3 ± 7.8

In the StructSeg dataset, the best localization performance was observed for Mandible R with an IOU of 0.805, while Opt Chiasma, being a relatively small organ, showed the lowest IOU of 0.104.

In summary, MedLAM exhibited reliable localization performance, particularly for larger organs. For smaller organs, despite lower IOU scores, the subsequent preprocessing step—expanding the 3D bounding box—ensured that these organs were adequately captured for segmentation, thereby ensuring practical applicability across a wide range of organ sizes. This methodology thus provides a practical solution to the challenge of localizing smaller organs with MedLAM, balancing out the performance across different organ sizes.

Evaluation of MedLSAM After validating the localization performance of MedLAM, we proceeded to examine the segmentation performance of the proposed MedLSAM framework. In our experiments, the Dice Similarity Coefficient (DSC) was employed as a measure to gauge the accuracy of our method. MedL-

Table 3: DSC (mean±std %) evaluation of 3D head-and-neck organs segmentation in the WORD dataset. The table compares the performance of SAM and MedSAM as segmentation basis within the MedLSAM framework, along with results from manually assisted localizations.

Localization	MedLAM		Manual	
Organs	SAM	MedSAM	SAM	MedSAM
Liver	55.5 ± 9.5	**67.1 ± 8.7**	**84.5 ± 6.3**	76.3 ± 6.7
Spleen	**60.9 ± 12.4**	40.3 ± 20.8	**87.2 ± 7.3**	60.2 ± 19.4
Kidney (L)	**78.5 ± 12.1**	66.2 ± 13.1	**92.0 ± 4.4**	72.3 ± 7.3
Kidney (R)	**83.0 ± 11.0**	62.3 ± 9.7	**92.9 ± 2.3**	67.1 ± 6.7
Stomach	**43.0 ± 13.1**	35.2 ± 15.9	**79.4 ± 8.6**	62.2 ± 14.6
Gallbladder	**33.2 ± 25.0**	27.8 ± 22.7	**72.9 ± 9.7**	67.3 ± 10.5
Esophagus	**24.6 ± 11.2**	22.1 ± 13.8	**68.2 ± 6.8**	48.2 ± 13.5
Pancreas	**34.1 ± 12.1**	28.4 ± 11.0	**64.0 ± 11.8**	56.7 ± 9.2
Duodenum	**22.9 ± 13.2**	18.5 ± 9.8	**59.7 ± 11.9**	42.3 ± 11.3
Colon	**18.5 ± 6.3**	12.8 ± 8.3	**42.9 ± 9.0**	22.0 ± 9.9
Intestine	**35.5 ± 9.3**	20.2 ± 8.8	**60.2 ± 7.2**	31.3 ± 9.1
Adrenal	**3.8 ± 5.2**	3.4 ± 3.1	**18.7 ± 12.1**	17.3 ± 10.0
Rectum	**37.6 ± 11.1**	29.9 ± 12.9	**74.4 ± 5.5**	53.7 ± 12.9
Bladder	**68.4 ± 21.3**	62.4 ± 17.8	**85.8 ± 11.6**	73.7 ± 10.5
Head of Femur (L)	**74.5 ± 7.7**	48.2 ± 15.6	**89.3 ± 4.7**	62.1 ± 12.8
Head of Femur (R)	**71.5 ± 5.3**	46.6 ± 10.1	**87.9 ± 3.6**	65.5 ± 7.0
Average	**45.9 ± 11.2**	37.7 ± 12.4	**72.5 ± 3.1**	54.9 ± 6.7

SAM utilizes the localization information from MedLAM, together with either SAM or MedSAM, to perform segmentation tasks.

In addition to the automatic localizations generated by MedLAM, we also included manual bounding boxes in our evaluation. These were simulated based on the ground-truth masks, following the same approach used for generating bounding boxes in MedSAM. During training, a bounding box prompt was generated from each ground-truth mask, with a random perturbation of 0-20 pixels introduced to mimic the potential inaccuracies that would be present in manually-drawn bounding boxes.

The DSC scores of StructSeg Task1 dataset are shown in Table 2, it is clear that MedLAM performed comparably to manual localisation in the context of small organs. For example, organs like the left and right eyes, MedLAM based SAM and MedSAM have DSC values of around 61-67%, which are close to the DSC values of 67-69% from manual localization. For minute organs such as the left and right lenses, MedLAM shows a similar DSC value to manual localisation (around 22% for SAM and 16% for MedSAM). However, MedLAM based SAM and MedSAM have lower performance in the context of the mandible (left and right), with DSC values of 48.3% and 43.5% for SAM and only 9.0% and 2.6% for MedSAM, compared to manual localization DSC values of 85.3% and 80.4%.

Meanwhile, our evaluation on the WORD dataset for abdominal organ segmentation, as shown in Table 3, revealed that MedLAM demonstrated robust

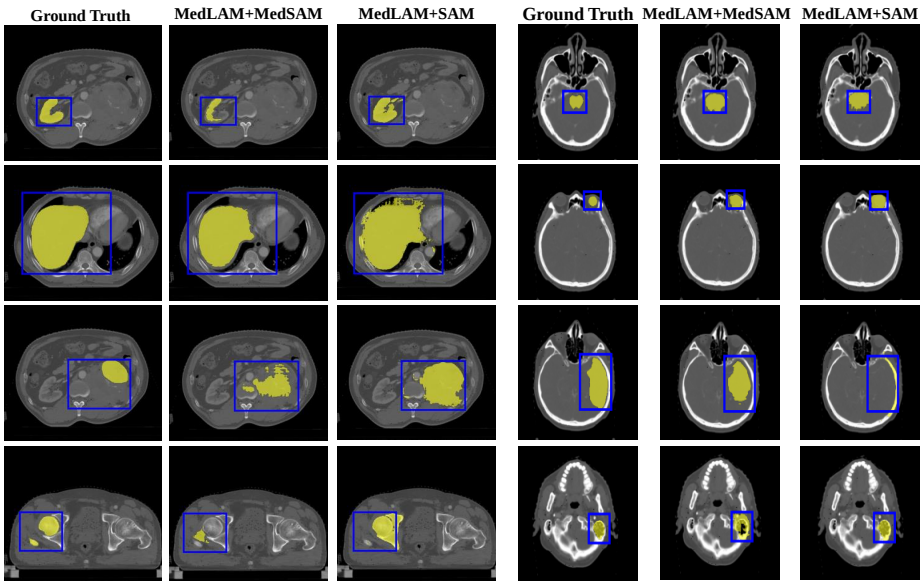

Fig. 6: Visualization examples of segmentation results on WORD and StructSeg Task1 datasets using pre-trained MedSAM and SAM, post landmark localization with MedLAM.

performance for organs such as the left and right kidneys. For instance, when integrated with SAM, the Dice Similarity Coefficient (DSC) reached 78.5% for the left kidney, and escalated to an impressive 83.0% for the right kidney. When combined with MedSAM, the model maintained satisfactory performance with a DSC of 66.2% for the left kidney and 62.3% for the right kidney. However, for smaller organs like the adrenal gland, both SAM and MedSAM yielded lower DSC values under 4% in comparison to a manual localization DSC of 18.7%.

4 Discussion & Conclusions

The presented work introduced MedLSAM, the first completely automated medical adaptation of the SAM model, designed to significantly alleviate the annotation workload in the segmentation of medical images. By cleverly integrating MedLAM, a few-shot localization framework, with SAM, the system was able to achieve comparable performance to SAM and its medical adaptations, yet required only minimal extreme point annotations for the entire dataset.

This endeavor was fueled by the observation that the spatial distributions of organs across different patients maintain strong similarities. Consequently, MedLAM was designed to project every part of the scans onto a shared 3D latent coordinate system, accurately localizing target anatomical parts within the body. Coupling this approach with SAM's segmentation capabilities led to an efficient and accurate process for image segmentation.

Moreover, MedLSAM demonstrated its effectiveness across two 3D datasets covering 38 different organs, providing robust evidence of its versatility. Importantly, this automated approach is not burdened by an increased annotation workload as data size increases. It also holds promise for direct integration with potential future 3D SAM models in the medical field, which could further enhance its performance and utility.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications 13(1), 4128 (2022)
2. Armato III, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al.: The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans. Medical physics 38(2), 915–931 (2011)
3. Bo, Z.H., Qiao, H., Tian, C., Guo, Y., Li, W., Liang, T., Li, D., Liao, D., Zeng, X., Mei, L., et al.: Toward human intervention-free clinical diagnosis of intracranial aneurysm via deep neural network. Patterns 2(2), 100197 (2021)
4. Chen, C.: Miccai 2022 mela challenge: Mediastinal lesion analysis (2022), `https://mela.grand-challenge.org/Dataset/`
5. Cheng, D., Qin, Z., Jiang, Z., Zhang, S., Lao, Q., Li, K.: Sam on medical images: A comprehensive study on three prompt modes. arXiv preprint arXiv:2305.00035 (2023)
6. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data 9(1), 601 (2022)
7. Grossberg, A.J., Mohamed, A.S., Elhalawani, H., Bennett, W.C., Smith, K.E., Nolan, T.S., Williams, B., Chamchod, S., Heukelom, J., Kantor, M.E., et al.: Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. Scientific data 5(1), 1–10 (2018)
8. He, S., Bao, R., Li, J., Grant, P.E., Ou, Y.: Accuracy of segment-anything model (sam) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 (2023)
9. He, S., Bao, R., Li, J., Stout, J., Bjornerud, A., Grant, P.E., Ou, Y.: Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets. arXiv preprint arXiv:2304.09324 (2023)
10. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis p. 101821 (2020)
11. Huang, Y., Yang, X., Liu, L., Zhou, H., Chang, A., Zhou, X., Chen, R., Yu, J., Chen, J., Chen, C., et al.: Segment anything model for medical images? arXiv preprint arXiv:2304.14660 (2023)
12. Ji, Y., Bai, H., Yang, J., Ge, C., Zhu, Y., Zhang, R., Li, Z., Zhang, L., Ma, W., Wan, X., et al.: Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. arXiv preprint arXiv:2206.08023 (2022)

13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
15. Kwan, J.Y.Y., Su, J., Huang, S.H., Ghoraie, L.S., Xu, W., Chan, B., Yip, K.W., Giuliani, M., Bayley, A., Kim, J., et al.: Radiomic biomarkers to refine risk models for distant metastasis in hpv-related oropharyngeal carcinoma. International Journal of Radiation Oncology* Biology* Physics 102(4), 1107–1116 (2018)
16. Lee, R.S., Gimenez, F., Hoogi, A., Miyake, K.K., Gorovoy, M., Rubin, D.L.: A curated mammography data set for use in computer-aided detection and diagnosis research. Scientific data 4(1), 1–9 (2017)
17. Lei, W., Xu, W., Gu, R., Fu, H., Zhang, S., Zhang, S., Wang, G.: Contrastive learning of relative position regression for one-shot object localization in 3d medical images. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 155–165. Springer (2021)
18. Lowe, V.J., Duan, F., Subramaniam, R.M., Sicks, J.D., Romanoff, J., Bartel, T., Yu, J.Q.M., Nussenbaum, B., Richmon, J., Arnold, C.D., et al.: Multicenter trial of [18f] fluorodeoxyglucose positron emission tomography/computed tomography staging of head and neck cancer and negative predictive value and surgical impact in the n0 neck: results from acrin 6685. Journal of Clinical Oncology 37(20), 1704 (2019)
19. Luo, X., Liao, W., Xiao, J., Chen, J., Song, T., Zhang, X., Li, K., Metaxas, D.N., Wang, G., Zhang, S.: Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. Medical Image Analysis 82, 102642 (2022)
20. Ma, J., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023)
21. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., et al.: Fast and low-gpu-memory abdomen ct organ segmentation: the flare challenge. Medical Image Analysis 82, 102616 (2022)
22. Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. arXiv preprint arXiv:2304.10517 (2023)
23. Revel, M.P., Boussouar, S., de Margerie-Mellon, C., Saab, I., Lapotre, T., Mompoint, D., Chassagnon, G., Milon, A., Lederlin, M., Bennani, S., et al.: Study of thoracic ct in covid-19: the stoic project. Radiology 301(1), E361–E370 (2021)
24. Vallieres, M., Kay-Rivest, E., Perrin, L.J., Liem, X., Furstoss, C., Aerts, H.J., Khaouam, N., Nguyen-Tan, P.F., Wang, C.S., Sultanem, K., et al.: Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific reports 7(1), 10117 (2017)
25. Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y., Arbel, T.: Medical sam adapter: Adapting segment anything model for medical image segmentation. arXiv preprint arXiv:2304.12620 (2023)
26. Yan, K., Cai, J., Jin, D., Miao, S., Guo, D., Harrison, A.P., Tang, Y., Xiao, J., Lu, J., Lu, L.: Sam: Self-supervised learning of pixel-wise anatomical embeddings in radiological images. IEEE Transactions on Medical Imaging 41(10), 2658–2669 (2022)

27. Yao, Q., Quan, Q., Xiao, L., Kevin Zhou, S.: One-shot medical landmark detection. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II 24. pp. 177–188. Springer (2021)
28. Zhang, K., Liu, D.: Customized segment anything model for medical image segmentation. arXiv preprint arXiv:2304.13785 (2023)
29. Zhang, Y., Jiao, R.: How segment anything model (sam) boost medical image segmentation? arXiv preprint arXiv:2305.03678 (2023)