ISSN 2464-4617 (print)
ISSN 2464-4625 (online)
Computer Science Research Notes - CSRN
http://www.wscg.eu
WSCG 2025 Proceedings

Efficient Semi-automatic Segmentation-labeling of Any Volumetric Medical Image

Jonas Kordt[1]
jonas.kordt@student.hpi.de

Sumit Shekhar[2]
sumit.shekhar@snu.edu.in

Christoph Lippert[1]
christoph.lippert@hpi.de

[1]Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam,
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany.

[2]Department of Computer Science and Engineering,
Shiv Nadar Institution of Eminence, Deemed to be University, Delhi NCR, India.

ABSTRACT

Regions of interest are often labeled in volumetric medical images either for research purposes or for diagnosis and treatment planning. However, labeling such segments manually is time-consuming and requires medical expertise, which makes it expensive. We design a novel semi-automatic 3D workflow which allows efficient segmentation-labeling of volumetric images. To this end, for a given 3D image we first manually label a subset of its 2D slices using MedSAM (a foundational model for segmenting any 2D medical image) via bounding-box prompting. Subsequently, we interpolate user-provided prompts for the remaining slices to automatically generate labels for them. This way, users can process a complete volumetric image while working on only a subset of its slices. We evaluate our method on the diverse set of medical image datasets from the Medical Segmentation Decathlon challenge. Our approach significantly reduces the labeling effort, around 67%, while only marginally reducing the segmentation accuracy compared to applying MedSAM slice-by-slice. Breaking out of the time-consuming slice-by-slice workflow with only a minor reduction in accuracy is a significant step in streamlining the process of semi-automatic labeling.

Keywords

medical-image labeling, segment anything, volumetric medical imaging, semi-automatic segmentation

1 INTRODUCTION

Segmentation of volumetric medical images, such as Magnetic Resonance Imaging (MRI) and Computer-aided Tomography (CT), facilitates quantitative analysis for disease diagnosis, pathological assessment, and surgical planning. In recent years, supervised deep learning techniques have proven quite successful for segmenting such volumetric images [17]. However, in order to achieve high-level of accuracy needed in the medical imaging domain these techniques demand sufficient amount of labeled training data.

Manual segmentation labeling of volumetric medical images is a tedious process, which requires the expertise of a trained medical practitioner making it expensive. Further, while manual labeling is the gold standard in terms of accuracy it also introduces inter- and intra-observer variability, leading to inconsistencies in the data analysis [14, 3]. In comparison, semi-automatic segmentation labeling combines the precision and control of manual approach with the speed and consistency of automated techniques [13]. However, due to high variability in volumetric medical images [1], it can be difficult for semi-automatic labeling tools to generalize well across different segmentation tasks. Moreover, even though such tools can go beyond slice-by-slice processing of volumetric images [10], thereby reducing the required manual effort, the effort-to-accuracy trade-off can still be improved significantly.

In this work, we address these challenges and propose a semi-automatic approach which: *(i)* generalizes well across a wide range of medical images, *(ii)* requires significantly less effort by breaking away from the slice-by-slice workflow and *(iii)* maintains high labeling accuracy. To this end, we leverage the recently introduced foundational model for the task of 2D medical image segmentation, namely MedSAM [11]. It is in turn a fine-tuned version of *Segment Anything Model* (SAM) [9], which shows surprising zero-shot performance for segmenting natural images. However, as SAM is trained on natural images it tends to struggle with medical images where boundaries are often softer in comparison to natural images [11]. While MedSAM achieves relatively high accuracy on 2D medical images, it still requires a slice-by-slice workflow for segmentation-labeling of 3D volumetric images where the expert draws a bounding box (prompt for MedSAM) around the region of interest on every single slice.

We propose a workflow that breaks this tedious slice-by-slice labeling procedure and requires significantly less effort from the experts, refereed to as "*3D workflow*" in the rest of the paper. We achieve this via novel prompt engineering strategies based on bounding box interpolation. The feasibility of such a workflow is shown by achieving accuracy comparable to state-of-the-art segmentation approaches. Further, we maintain best effort-to-accuracy trade-off for a variety of medical image datasets used in the Medical Segmentation Decathlon challenge [1]. In particular, the bilinear interpolation showed the ability to reduce the manual prompting input to just 33% with only a minor reduction in dice-score (DSC) of 0.0175 on average. For some datasets, including brain tumor, heart, and liver segmentation, it is even possible to reduce manual prompting to as low as 9% with only a DSC reduction of around 0.01.

2 RELATED WORK

In the field of volumetric image segmentation there are three overarching approaches. Namely manual segmentation, semi-automatic segmentation, and automatic segmentation.

2.1 Manual and Semi-automatic Segmentation

Many different manual and semi-automatic tools are usually combined in a single segmentation application. The landscape of segmentation applications includes free-to-use software, and commercial software. Among free-to-use software, options include ITK-SNAP[1] [21], MITK[2] [18], 3D-Slicer[3] [4], and VISIAN[4] [10]. A commercial alternative is Encord[5] [6]. Popular manual segmentation tools that these applications include are simple pixel brushes and outlining tools. Semi-automatic segmentation tools, which reduce manual effort, combine user input with various automated algorithms, such as region-growing, thresholding, dilation and erosion, or even more complex learning-based approaches.

2.2 Promptable Segmentation Using Deep Learning

Recently, the Segment Anything Model (SAM) introduced by Meta has emerged as a powerful segmentation approach for images in wild. However, despite good zero-shot performance on natural images, multiple studies have found the zero-shot performance on medical images to be generally lower than state of the art deep learning models and varying a lot depending on the dataset [12, 7, 15]. To improve SAM's performance on medical images, various fine-tuning and adaptation approaches have been proposed. MedSAM [11] fine-tunes SAM's image encoder and prompt decoder using over 1 million medical image masks and applies specific preprocessing for different image types. SAM-Med2D [2], an enhanced version of SAM with adapter layers, is fine-tuned with a very large dataset of 2D medical images and segmentation masks [20], yet shows lower dice scores than other methods. SAM-Med3D [16] extensively modifies SAM for 3D images, trained from scratch for improved segmentation quality, without comparison to advanced models. 3DSAM-adapter [5], tailored for 3D images, reuses and adapts SAM's components, showing superior performance on most datasets compared to SAM and other benchmarks. The idea of promptable segmentation has also been transferred to other deep learning approaches which are not based on SAM. One example is the One-Prompt Segmentation [19], which combines the strengths of one-shot methods and prompting. Additionally, Ma et al. [11] suggest a promptable version of nnU-Net [8] and use it as a benchmark for MedSAM. The promptable nnU-Net works by encoding a bounding box prompt in a binary bask and using this mask as a second image channel during training of and inferencing with the nnU-Net. We discuss how this promptable version of nnU-Net differs from SAM and MedSAM in the supplementary material.

3 METHOD

3.1 Segmentation-Labeling of Any Volumetric Medical Image

Segment Anything Model (SAM): It is a foundation model for promptable segmentation which is trained on more than a billion segmentation masks for 2D natural images. It consists of three main parts: the image encoder, the prompt encoder, and the mask decoder. The image encoder is a pre-trained ViT (vision transformer) which generates the image embedding. The generation of the embedding only has to be done once per image and can then be continuously reused for multiple prompts. The prompt encoder process input prompts provided in the form of: points (foreground and background), a bounding box, text, or a preexisting segmentation mask to be refined. To compute the resulting segmentation mask, the mask decoder interprets the image embedding (with the optionally included preexisting mask embedding) and the prompt encodings. The resource-intensive image encoding task requires a high performance GPU in order to run efficiently. On the other hand, the lightweight prompt encoder and mask decoder can run even on a web browser

[1] http://www.itksnap.org/
[2] https://www.mitk.org/
[3] https://www.slicer.org
[4] https://visian.org.
[5] https://encord.com/

using a consumer-grade CPU. It implies that given a GPU server, users can interactively use SAM on a web browser via consumer-grade machines.

Fine-tuning and Adaptation for Medical Images:

Since SAM struggles with medical images [12, 7, 15] various approaches for fine-tuning and adaptation have been proposed. One popular fine-tuning approach is MedSAM [11], for which the image encoder and mask decoder were fine-tuned using over 1.5 million medical image segmentation masks encompassing a variety of medical imaging scenarios. The prompt encoder remains unchanged from SAM. The fine-tuning dataset includes both 2D and 3D medical images and different kinds of structures, such as organs and tumors. During fine-tuning, only bounding box prompts were used. In addition to the fine-tuned model, MedSAM also uses specific pre-processing for the medical images.

3D Workflow for Segmentation-Labeling:

A typical labeling workflow for applying a 2D promptable model, such as MedSAM (or SAM), to volumetric medical images is by operating on a slice-by-slice basis. In the proposed 3D workflow, the user should no longer have to provide a prompt for every single slice of the volumetric image. Instead, from sparse user-provided prompts we estimate the rest of the prompts via prompt-interpolation strategies. Further, since bounding box prompts generally achieve better results than point prompts [12, 15, 11], we only consider that and leave experimentation with other prompts for future work. An overview of our workflow is depicted in Fig. 1. Moreover, our workflow can be implemented in a client-server fashion wherein clients/users can access the system via a web browser, see Fig. 2.

3.2 Prompt Engineering

Nearest-neighbor Interpolation:

A straightforward way to estimate additional bounding box prompts is via extending the same user-provided bounding box for a given slice to its n neighbors. By propagating the prompt to 1 neighboring slice (on both sides), we already reduce the prompting-effort by 66% since a single prompt is now used for 3 slices instead of just 1 slice. However, we can extend this idea even further by increasing the value of n. For example, for $n = 5$, a single prompt is used for 11 slices, reducing the prompting-effort by more than 90%. Obviously, this comes with a trade-off wherein the further we extend a user-provided prompt the greater is the inaccuracy of the output labels for the propagated prompts.

Bilinear Interpolation:

This limitation can be mitigated to an extent by linearly interpolating bounding box boundaries instead of naively using the same bounding box for the neighboring slices. To this end, two bounding boxes at the edges of a slice-interval are provided by the user. The bounding boxes between the two inputs are then generated in the following manner. A bounding box B_s on a slice with index $s \in \mathbb{N}$ is defined by two points (x_s^{min}, y_s^{min}) and (x_s^{max}, y_s^{max}). Let the interpolation interval go from slice $a \in \mathbb{N}$ to slice $b \in \mathbb{N}$. The bounding boxes B_a and B_b on slices a and b are provided by the user. For a slice with index $i \in \mathbb{N}$ with $a < i < b$, we calculate x_i^{max} as:

$$x_i^{max} = \left(\frac{i-a}{b-a}\right) \cdot x_b^{max} + \left(1 - \frac{i-a}{b-a}\right) \cdot x_a^{max} \qquad (1)$$

y_i^{max}, x_i^{min}, and y_i^{min} are calculated similarly.

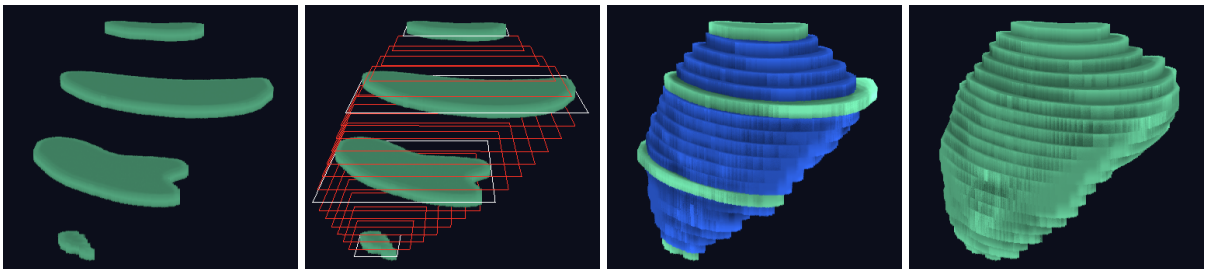
While the bilinearly interpolated prompts can better capture structures which are aligned diagonally to the slicing direction (see Fig. 3c), highly irregular structures are still hard to capture. Thus, the trade-off between reducing effort and the accuracy of bounding-box prompts still exists. Nevertheless, bilinear interpolation achieves better trade-off than the nearest-neighbor interpolation, especially for larger propagation distances (see Sec. 4.1).

4 EVALUATION

We compare our 3D workflow against the following approaches: SAM [9], MedSAM [11], and SAM-Med3D (and SAM-Med3D-turbo) [16]. Using SAM and MedSAM we simulate the normal slice-by-slice workflow, where the model is prompted with a bounding box on every single slice. SAM-Med3D and SAM-Med3D-turbo, on the other hand, constitute an alternate workflow for directly segmenting 3D volumetric images based on point prompts. For this, we provide ten 3D point prompts to the pre-trained models.

4.1 Experimental Setup

For our experiments we adopt the data pre-processing and normalization method which MedSAM was trained with [11]. To evaluate SAM, we use the most powerful ViT-H image encoder version, while MedSAM is based on the smaller ViT-B image encoder [11]. For both the nearest-neighbor and the bilinear interpolation techniques we simulate various amounts of user input. This corresponds to varying interpolation intervals and different propagation distances. Specifically, we simulate interpolation intervals of lengths 2, 4, 6, 8, and 10 slices (excluding the top and bottom slices



(a) Manual labels (b) New prompts (c) Suggested labels (d) Ground truth

Figure 1: Overview of our 3D workflow: (a) A subset of slices are labeled manually using MedSAM via bounding box prompts for segmenting out the structure of Spleen. (b) The initial user-provided prompts are interpolated to create new prompts for intermediate slices. (c) The generated prompts are used to create new labels for intermediate slices. Note, how the suggested labels are quite similar to the to the ground truth (d) for a smooth structure, such as the Spleen.

which are used as input for the interpolation) – for bilinear, and propagation distances of 1 to 5 slices (both above and below the input slice) – for nearest-neighbor. This results in approximately 33%, 20%, 14%, 11%, and 9% user prompts for both the interpolation strategies respectively. Additionally, to avoid interpolating between the usually small edge bounding boxes of a structure, we ensure that at least two interpolation intervals and at least two propagation inputs are used. For the SAM-Med3D and SAM-Med3D-turbo benchmarks, we use the available evaluation script, with 10 point prompts. The first point prompt is a random point from the foreground region. The following points are iteratively placed in a random position within the error region [16].

4.2 Dataset Description

In order to better match the common real world application scenario of semi-automatically segmenting one structure at a time while viewing one image modality at a time, we converted the Medical Segmentation Decathlon dataset [1] to this scenario as follows. As the brain tumor segmentations include different subclasses of the tumor, we combine them all to a single class. We use the FLAIR-MRI channel, as FLAIR is commonly used for tumor segmentation and produces good contrast, and disregard the other channels. The liver segmentation include the healthy liver and small tumors within the liver. We combine the classes to result in one class for the whole liver. The hippocampus segmentations are split into anterior and posterior hippocampus. We combine both classes to result in one segmentation for the whole hippocampus. Similarly, prostate segmentation are split into the peripheral zone and the transition zone. We again combine them to one class of the whole prostate. Additionally, we choose the T2-MRI channel and disregard the ADC-MRI channel. The pancreas segmentations include one class for the healthy pancreas and one class for small tumors. We combine the classes into one class of the whole pancreas. The hepatic tumor dataset includes one class for hepatic vessels and one class for tumors. As the hepatic vessels are many small disconnected structures which are entirely distinct from the tumor, we choose the tumor class and disregard the non-optimal vessel class. All other datasets already have a single channel and single foreground segmentation class and remain unchanged. Overall, even with our simplification to a single image channel and a single foreground segmentation class, the Medical Segmentation Decathlon datasets remain a diverse set of volumetric medical image segmentation tasks.

4.3 Quantitative Results

The quantitative results of our experiments are presented in Tab. 1. Note, that the results of both SAM and MedSAM are upper-bound baselines, since both methods are prompted with a ground truth bounding box. In contrast the proposed method uses bounding box interpolation to achieve similar accuracy. Some key observations are as follows. Firstly, the difference in using SAM vs. MedSAM is significant showing the efficacy of MedSAM for medical images. Secondly, the alternate 3D workflow using point prompts and the 3D models SAM-Med3D and SAM-Med3D-turbo, while arguably requiring less user effort, is not competitive with regards to accuracy. This also remains true when comparing it to our prompt engineering methods instead of the slice-by-slice workflow using SAM or MedSAM. Even when using only 9% of user prompts, both the nearest-neighbor- and bilinear- interpolation outperform SAM-Med3D and SAM-Med3D-turbo. The only exceptions are brain-tumor and heart segmentation.

Our interpolation-based approach is obviously less accurate than applying MedSAM slice-by-slice. However, the labeling quality, especially for 33% user prompts, is only slightly worse than the slice-by-slice approach using MedSAM and still outperforms the case of using SAM instead (see Tab. 1). For both our interpolation techniques we see a relatively even reduction in segmentation quality as the amount of user prompts is reduced. Between the two we see superior

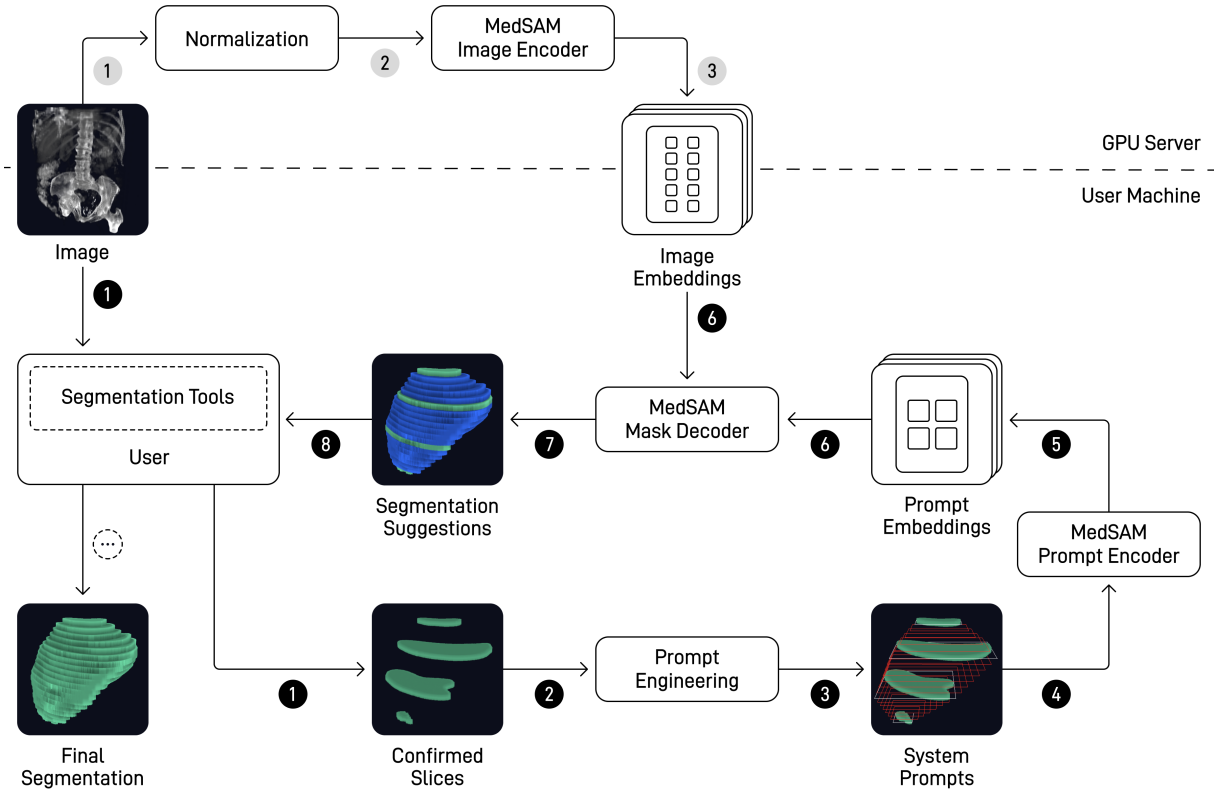
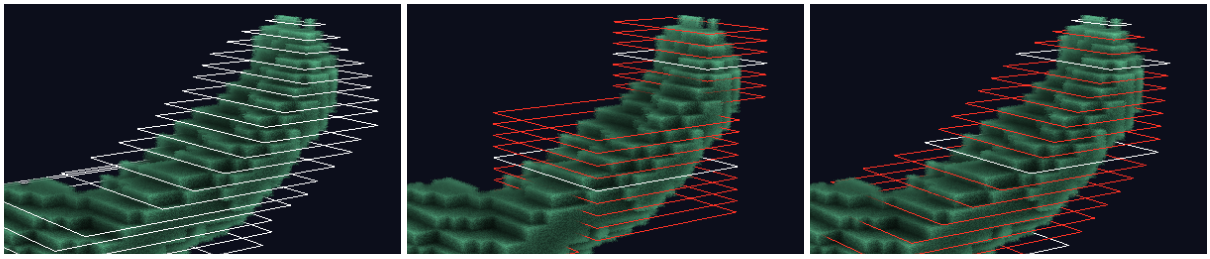


Figure 2: Flow diagram of the client-server architecture. The image is made accessible to both the user machine and the GPU server. The GPU server handles image normalization and generates image embeddings for user machine to download and cache (1 to 3). The segmentation labeling (❶ to ❽) starts when the user confirms segmentation slices (❶) using the segmentation tools. Confirmed slices are input for prompt engineering (❷) which creates system prompts (❸). These prompts are encoded and combined with corresponding image embeddings provided by the GPU server to generate new segmentation suggestions (❹ to ❼). The user checks and corrects these suggestions (❽) thereby obtaining the final (⋯) output.



(a) Ground truth boxes (b) Nearest Neighbor (c) Bilinear

Figure 3: Hippocampus volumes rendered in 3D with slice-wise bounding boxes. White bounding boxes are user-provided (simulated from ground truth) and red bounding boxes are estimated via (b) nearest-neighbor interpolation and (c) bilinear interpolation respectively.

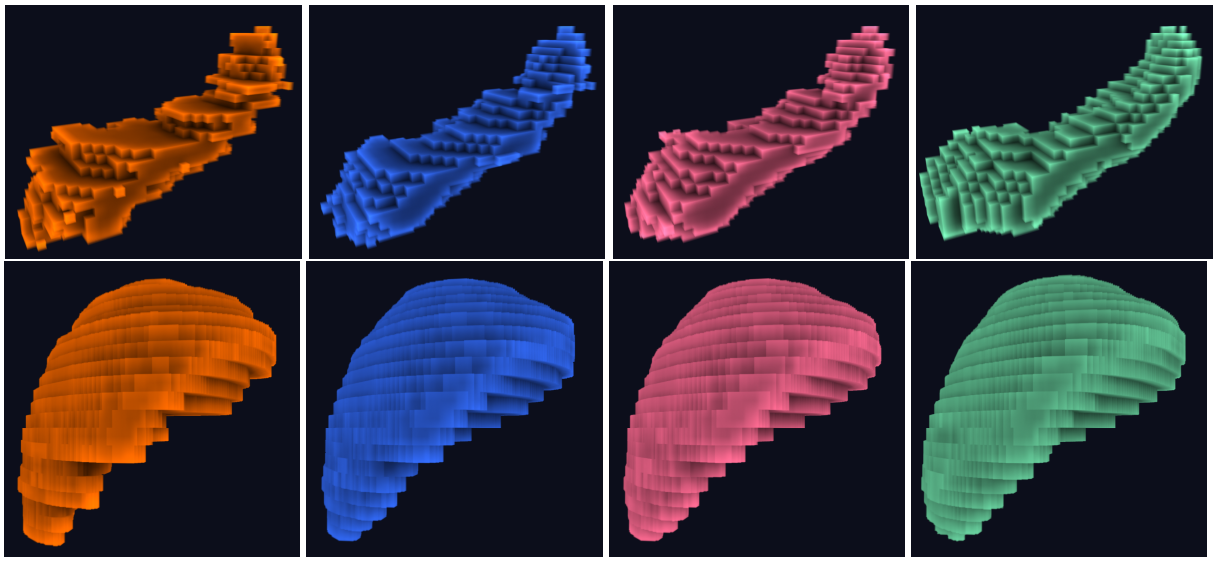
labeling quality for the bilinear interpolation, with an improvement in DSC of around 0.01.

4.4 Qualitative Results

To further qualitatively analyze the accuracy of nearest-neighbor- vs. bilinear- interpolation we present examples for hippocampus and spleen segmentation in Fig. 4. The shown cases are the same used for showing slice-wise bounding boxes in Fig. 3. First of all, we observe that the slice-by-slice application of MedSAM (see Sec. 4.3) is quite close to the ground truth segmentation (see Sec. 4.3). Thus, the result only requires minor manual clean up. The nearest-neighbor interpolation approach, as expected, results in edgy segmentation where the chunks of slices sharing the same bounding box prompts are clearly visible (see Sec. 4.3). These chunks become even more obvious when using a larger interpolation interval. In comparison, the bilinear interpolation results in smoother segmentation boundaries. However, for larger interpolation intervals the quality in this case also degrades significantly (see Fig. 5a).

Table 1: Dice-score (DSC) achieved by bilinear- and nearest-neighbor- interpolation of MedSAM [11] prompts compared to applying SAM [9] and MedSAM slice-by-slice (using ground-truth bounding box prompts), and SAM-Med3D and SAM-Med3D-turbo [16] prompted with ten 3D point prompts. All applied to the Medical Segmentation Decathlon dataset [1]. For both the interpolation techniques, user input was simulated on a varying percentage of slices. The best approach in terms of accuracy is marked in green and the next best in blue.

	Brain T.	Heart	Liver	Hippoca.	Prostate	Lung T.	Pancreas	Hepatic T.	Spleen	Colon T.	**Average**
MedSAM	0.830	0.846	0.955	0.822	0.903	0.792	0.822	0.7820	0.949	0.783	**0.848**
SAM	0.756	0.839	0.907	0.778	0.903	0.804	0.749	0.764	0.935	0.735	**0.817**
SAM-Med3D-turbo	0.863	0.878	0.876	0.118	0.661	0.739	0.659	0.492	0.817	0.508	**0.661**
SAM-Med3D	0.760	0.692	0.849	0.357	0.434	0.470	0.365	0.494	0.759	0.380	**0.556**
Nearest-neighbor											
33% user prompts	0.828	0.843	0.953	0.782	0.874	0.769	0.799	0.745	0.929	0.724	**0.825**
20% user prompts	0.824	0.839	0.949	0.728	0.840	0.759	0.771	0.727	0.901	0.701	**0.804**
14% user prompts	0.817	0.833	0.945	0.676	0.814	0.750	0.745	0.716	0.871	0.692	**0.786**
11% user prompts	0.812	0.826	0.941	0.618	0.793	0.743	0.718	0.711	0.854	0.689	**0.770**
9% user prompts	0.804	0.822	0.935	0.586	0.785	0.741	0.698	0.706	0.830	0.688	**0.759**
Bilinear											
33% user prompts	0.830	0.843	0.955	0.808	0.870	0.781	0.804	0.752	0.938	0.729	**0.831**
20% user prompts	0.829	0.842	0.953	0.790	0.836	0.777	0.774	0.733	0.917	0.707	**0.816**
14% user prompts	0.827	0.841	0.951	0.770	0.783	0.768	0.742	0.716	0.886	0.684	**0.797**
11% user prompts	0.823	0.840	0.948	0.749	0.747	0.769	0.708	0.704	0.860	0.679	**0.783**
9% user prompts	0.820	0.838	0.944	0.717	0.738	0.764	0.670	0.696	0.827	0.678	**0.769**



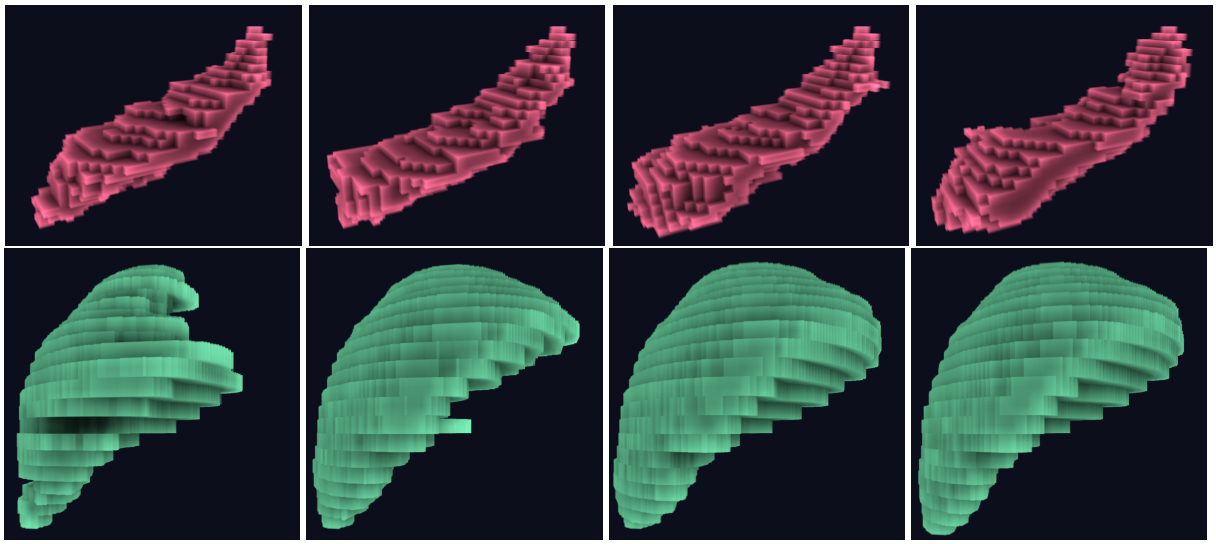
(a) Nearest-nbr. (b) Bilinear (c) MedSAM on all (d) Ground truth

Figure 4: Hippocampus and Spleen segmentations generated with different methods, all using MedSAM, compared to ground truth. In all images the individual voxels are quite visible due to low resolution (1 voxel per *mm*) compared to the size of the hippocampus ($\sim 37mm$ in length). For both nearest-neighbor and bilinear interpolation case we use only 33% of groud-truth bounding boxes. Note the (disconnected) chunks resulting from nearest-neighbor interpolation and the loss of fine details but preservation of overall shape resulting from bilinear interpolation.

5 DISCUSSION

We make use of MedSAM based per-slice approach along with novel prompt engineering for our 3D workflow. Alternatively, one can directly apply 3D based segment anything model like SAM-Med3D. However, evaluating it on existing datasets reveals a notable deficiency in segmentation quality, likely caused by the small amount of user input given to the model. Although the further fine-tuned version, SAM-Med3D-turbo, shows some improvement, it fails to compete effectively in more than two datasets, specifically the brain tumor and heart datasets. This highlights the critical balance between minimizing user input and maintaining segmentation quality. It remains to be seen if further fine-tuning or architectural enhancements could make these 3D models more competitive.

In contrast, our approach, leveraging MedSAM coupled with the presented prompt engineering methods, strikes a more advantageous balance. However, fine details can sometimes be lost, a limitation not unique to our approach but inherent to MedSAM, even in its slice-by-slice application. Improving MedSAM or potentially replacing it with another 2D model, such as SAM-Med2D [20], could be viable solutions.



(a) 9% (b) 14% (c) 33% (d) 100%

Figure 5: Visualization of how the segmentation-labeling quality improves for Hippocampus and Spleen as the amount of ground-truth bounding box is increased from 9% to 33% in case of bilinear interpolation. Note how similar are the results using just 33% ground-truth boxes in comparison to 100% especially for Spleen.

While we have presented an extensive evaluation of the proposed prompt engineering methods, the 3D workflow as such has not been tested with actual users. However, with the presented prompt engineering methods we have laid out the necessary foundation for implementing such a workflow in any medical image segmentation application. Further, our linear interpolation strategy might not perform well for structures changing abruptly across slices.

6 CONCLUSION

We present a semi-automatic segmentation-labeling approach for volumetric images that works for a wide variety of medical imaging scenarios and breaks away from the traditional slice-by-slice labeling workflow. Our approach, is based on promptable segmentation for 2D medical images using MedSAM. To this end, we design a 3D workflow that requires significantly less effort from experts compared to manually prompting MedSAM on every slice. As part of our design choice we evaluate two different prompt propagation strategies namely nearest-neighbor- and bilinear-interpolation. In particular, the bilinear interpolation showed the ability to reduce the manual prompting input to 33% with only a minor reduction in DSC (0.0175) on average. The early stage of SAM-based models for medical image segmentation leaves room for future improvement, and the subjective evaluation of the proposed 3D workflow remains as future work. Furthermore, our approach can benefit from integrating learning-based adaptive interpolation techniques.

REFERENCES

[1] Michela Antonelli et al. “The Medical Segmentation Decathlon”. In: *Nature Communications* 13.1 (July 2022). DOI: 10.1038/s41467-022-30695-9. URL: https://doi.org/10.1038/s41467-022-30695-9.

[2] Junlong Cheng et al. *SAM-Med2D*. 2023. DOI: 10.48550/ARXIV.2308.16184. URL: https://arxiv.org/abs/2308.16184.

[3] Elise C. Covert et al. “Intra- and inter-operator variability in MRI-based manual segmentation of HCC lesions and its impact on dosimetry”. In: *EJNMMI Physics* 9.1 (Dec. 2022). ISSN: 2197-7364. DOI: 10.1186/s40658-022-00515-6. URL: http://dx.doi.org/10.1186/s40658-022-00515-6.

[4] Andriy Fedorov et al. “3D Slicer as an image computing platform for the Quantitative Imaging Network”. In: *Magnetic Resonance Imaging* 30.9 (Nov. 2012), pp. 1323–1341. DOI: 10.1016/j.mri.2012.05.001. URL: https://doi.org/10.1016/j.mri.2012.05.001.

[5] Shizhan Gong et al. *3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation*. 2023. DOI: 10.48550/ARXIV.2306.13465. URL: https://arxiv.org/abs/2306.13465.

[6] Ulrik Stig Hansen et al. “Novel artificial intelligence-driven software significantly shortens the time required for annotation in

computer vision projects". In: *Endoscopy International Open* 09.04 (Apr. 2021), E621–E626. DOI: 10.1055/a-1341-0689. URL: https://doi.org/10.1055/a-1341-0689.

[7] Sheng He et al. *Computer-Vision Benchmark Segment-Anything Model (SAM) in Medical Images: Accuracy in 12 Datasets*. 2023. DOI: 10.48550/ARXIV.2304.09324. URL: https://arxiv.org/abs/2304.09324.

[8] Fabian Isensee et al. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation". In: *Nature Methods* 18.2 (Dec. 2020), pp. 203–211. DOI: 10.1038/s41592-020-01008-z. URL: https://doi.org/10.1038/s41592-020-01008-z.

[9] Alexander Kirillov et al. *Segment Anything*. 2023. DOI: 10.48550/ARXIV.2304.02643. URL: https://arxiv.org/abs/2304.02643.

[10] Jonas Kordt et al. "Interactive Volumetric Region Growing for Brain Tumor Segmentation on MRI using WebGL". In: *The 26th International Conference on 3D Web Technology*. ACM, Nov. 2021. DOI: 10.1145/3485444.3487640. URL: https://doi.org/10.1145/3485444.3487640.

[11] Jun Ma et al. "Segment anything in medical images". In: *Nature Communications* 15.1 (2024), p. 654. ISSN: 2041-1723. DOI: 10.1038/s41467-024-44824-z.

[12] Maciej A. Mazurowski et al. "Segment anything model for medical image analysis: An experimental study". In: *Medical Image Analysis* 89 (Oct. 2023), p. 102918. ISSN: 1361-8415. DOI: 10.1016/j.media.2023.102918. URL: http://dx.doi.org/10.1016/j.media.2023.102918.

[13] K.K.D. Ramesh et al. "A Review of Medical Image Segmentation Algorithms". In: *EAI Endorsed Transactions on Pervasive Health and Technology* (July 2018), p. 169184. DOI: 10.4108/eai.12-4-2021.169184. URL: https://doi.org/10.4108/eai.12-4-2021.169184.

[14] Félix Renard et al. "Variability and reproducibility in deep learning for medical image segmentation". In: *Scientific Reports* 10.1 (Aug. 2020). DOI: 10.1038/s41598-020-69920-0. URL: https://doi.org/10.1038/s41598-020-69920-0.

[15] Saikat Roy et al. *SAM.MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model*. 2023. DOI: 10.48550/ARXIV.2304.05396. URL: https://arxiv.org/abs/2304.05396.

[16] Haoyu Wang et al. *SAM-Med3D*. 2023. DOI: 10.48550/ARXIV.2310.15161. URL: https://arxiv.org/abs/2310.15161.

[17] Risheng Wang et al. "Medical image segmentation using deep learning: A survey". In: *IET Image Processing* 16.5 (2022), pp. 1243–1267. DOI: https://doi.org/10.1049/ipr2.12419.

[18] Ivo Wolf et al. "The medical imaging interaction toolkit (MITK): a toolkit facilitating the creation of interactive software by extending VTK and ITK". In: *SPIE Proceedings*. Ed. by Jr. Robert L. Galloway. SPIE, May 2004. DOI: 10.1117/12.535112. URL: https://doi.org/10.1117/12.535112.

[19] Junde Wu and Min Xu. *One-Prompt to Segment All Medical Images*. 2023. DOI: 10.48550/ARXIV.2305.10300. URL: https://arxiv.org/abs/2305.10300.

[20] Jin Ye et al. *SA-Med2D-20M Dataset: Segment Anything in 2D Medical Imaging with 20 Million masks*. 2023. DOI: 10.48550/ARXIV.2311.11969. URL: https://arxiv.org/abs/2311.11969.

[21] Paul A. Yushkevich et al. "User-guided 3D active contour segmentation of anatomical structures: Significantly improved efficiency and reliability". In: *NeuroImage* 31.3 (2006), pp. 1116–1128. ISSN: 1053-8119. DOI: https://doi.org/10.1016/j.neuroimage.2006.01.015. URL: https://www.sciencedirect.com/science/article/pii/S1053811906000632.