Adapt While You Annotate: The Missing Complement to Interactive Segmentation

Parhom Esmaeili[1], Chayanin Tangwiriyasakul[1], Eli Gibson[2], Sebastien Ourselin[1], and M. Jorge Cardoso[1]

[1] School of Biomedical Engineering and Imaging Sciences, KCL, UK
[2] Siemens Healthineers, Princeton NJ, USA
parhom.esmaeili@kcl.ac.uk

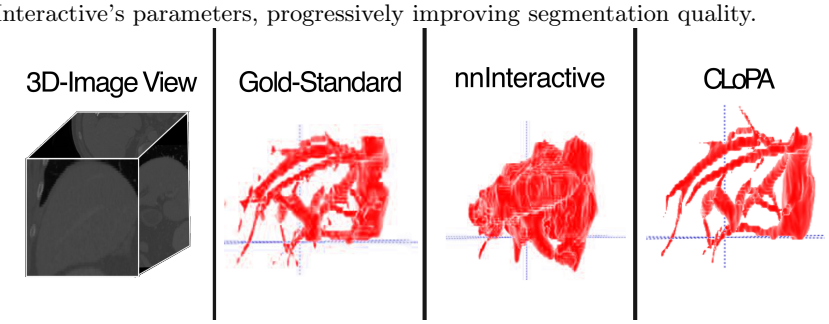
Abstract. Interactive segmentation enables clinicians to guide annotation, but existing zero-shot models like nnInteractive fail to consistently reach expert-level performance across diverse medical imaging tasks. Because annotation campaigns produce a growing stream of task-specific labelled data, online adaptation of the segmentation model is a natural complement to zero-shot inference. We demonstrate this using a continual adaptation strategy that does not introduce any new parameters or changes to the inference pipeline. We only tune a small fraction of a baseline model's parameters (the instance normalisation parameters) on the annotation cache triggered by lightweight episode scheduling. We elevate the performance of a state-of-the-art baseline model on eight Medical Segmentation Decathlon tasks spanning diverse anatomical targets and imaging characteristics to expert-level performance, even for tasks where the baseline model previously failed, with the majority of gains realised after a single training episode. We show that the benefits of tuning also depends on task characteristics, with saturating performance gains in targets with complex geometries (e.g., hepatic vessels), ambiguous boundaries (brain tumour core), or where there is a mismatch between the spatial scale of pretrained features and the scale of the target (hippocampus). These all suggest a need for feature-representation alignment in the most challenging scenarios.

Keywords: Interactive segmentation, instance incremental learning, medical image analysis, dataset annotation, parameter-efficient fine-tuning

1 Introduction

Large-scale annotated datasets in medical imaging are bottlenecked by data-sharing restrictions [10] and the cost of manual segmentation [11]. Domain shifts across medical centres further limit the applicability of static pre-trained segmentation models like nnU-Net [6] for annotation. Interactive segmentation addresses this by letting clinicians guide the process through prompts such as clicks, scribbles, or bounding boxes [7]. Among recent models, nnInteractive [7] achieves strong zero-shot generalisation across diverse anatomies. Yet even nnInteractive cannot consistently reach expert-level performance with low-effort click

Fig. 1. Overview of CLoPA: As annotated samples are produced, the annotation cache grows and periodically triggers training episodes that fine-tune only a small subset of nnInteractive's parameters, progressively improving segmentation quality.



prompts across all tasks [5], making zero-shot models impractical for large-scale annotation campaigns where both speed and reliability are essential.

Since annotation is inherently repetitive, annotated samples produced during a campaign constitute a growing task-specific dataset that can be exploited for online model adaptation. This motivates an instance-incremental learning approach: rather than relying solely on the zero-shot model, we progressively fine-tune it on the annotation stream to close the gap to expert-level performance. Crucially, adaptation must remain lightweight—both to avoid overfitting on the small, growing cache and to preserve the strong initialisation from pre-training.

In this work, we make the following contributions: (1) we demonstrate a proof-of-concept (CLoPA) to outline the effectiveness of a crude adaptation strategy that fine-tunes only a small subset parameters of a base foundation model (e.g., nnInteractive) during the annotation workflow, requiring no new parameters and no changes to inference; (2) we demonstrate across eight MSD tasks that this lightweight adaptation rapidly achieves expert-level performance, including on tasks where the base model previously failed; and (3) we extend the interactive segmentation evaluation protocol of Esmaeili et al. [5] with trajectory metrics that capture adaptation dynamics over time.

1.1 Related Work

Interactive segmentation algorithms incorporate image data and user guided prompts (e.g., clicks, scribbles, bounding boxes or lassos [7]) to generate output segmentations on prompted targets. Current state-of-the-art methods employ deep learning to design these algorithms; and during training networks are typically trained with a synthetic user-in-the-loop to iteratively prompt from the error-region between a reference annotation and a prediction [3, 4, 14, 12, 16, 7]. Zero-shot models, intended for general out-of-the-box use, generate training examples by pooling huge quantities of different segmentation datasets [14, 7, 3, 4] to sample image-annotation pairs. Training data diversity is introduced by pseudo-labelling strategies: but classical methods like SLIC [1] produce coarse

parcellations, while SAM-based projection [9, 12] used by nnInteractive yields mostly blobby instances due to the domain gap from natural images [7].

Existing methods address prompt ambiguity with multiple candidate masks [9, 12], composite training annotations [7], or more restrictive prompts such as bounding boxes and lassos [4, 7]. However, restrictive prompts are laborious for complex geometries, and no current method reliably achieves expert level performance using only low-effort click interactions [5]. Additionally, fixed patch sizes and zooming heuristics [4, 7] used for volumetric inference are suboptimal for targets with low volume fractions (sparse, branching structures), where CNN based methods particularly struggle [7].

In summary, zero-shot models lack the inductive biases to consistently reach expert-level performance, making them unsuited for large-scale annotation. Since annotation is inherently repetitive, adaptation to the task context is a natural complement. In-context learning approaches [15] plateau below specialist performance and do not scale with dataset size. Continual parameter tuning has been explored in 2D [17] interactive segmentation, but to our knowledge not in 3D.

2 Methodology

We consider a practical data-annotation scenario: a single, fixed binary segmentation task where annotated samples are cached as they are produced. In this work, we adapt a base interactive-segmentation foundation model, namely nnInteractive [7], on a stream of incoming annotations using full memory replay, without modifying the inference pipeline. Our simple approach has two simple components: (1) *when* to trigger adaptation (training episodes), and (2) *how* to adapt. Our training episode is only triggered once the annotation cache obtains a bank of samples unassigned to a training split. The simple criterion we use is that at least 5 unassigned samples are required (from $k_{\mathcal{M}} = 0.2$, i.e., $\geq 1/k_{\mathcal{M}}$ samples to allow for a validation split) and a quantity of unassigned samples equivalent to 25% of the dataset ($k_{\mathcal{D}} = 0.25$) are cached.

2.1 Training Configurations

Tunable parameters: We freeze all pretrained weights bar one configuration in this proof-of-concept demonstration: the instance normalisation [13] affine parameters (scale and bias). We denote this configuration **CLoPA-I.N** in all tables. Instance normalisation modulates per-channel feature statistics independently for each sample, making it a natural fit for task-specific style and contrast adaptation without altering the learned spatial filters. Because these parameters constitute a tiny fraction of the total model ($<0.01\%$), the risk of overfitting on the small annotation cache is minimal, while still providing capacity to recalibrate feature distributions for the target anatomy.

Data synthesis: Following the nnU-Net [6] data engine which underpins nnInteractive, we sample 192^3 patches uniformly across foreground and background classes from the annotation cache, applying nnInteractive's preprocessing and augmentation pipeline. For each gradient step, we simulate a click-based

interaction loop with one foreground and one background point sampled per interaction step. For initialisation these are sampled directly from the ground truth; for edits they are sampled from the false-negative error region.

We currently employ fine-tuning with only click-prompting as this is the least laborious form of user-interaction. For training we use an unweighted Dice Cross-Entropy loss [8] averaged across interaction time-steps, same as Wong et al. [16]. For a maximum number of interaction time-steps, N our loss is:

$$L = \frac{1}{N} \sum_{i=1}^{N} L_{Dice}(m_i, y) + L_{CE}(m_i, y) \tag{1}$$

Where m_i indicates the softmaxed prediction for the batch at interaction-step i. Each loss term is averaged across batch samples, if a batch sample reaches a termination condition (Dice Score of 1) then it is not accounted for on subsequent iteration time-steps in the loss calculation. For each gradient update we employ $N = 5$ interaction steps per gradient update and a fixed initial batch size of 2. For each training episode, an initial learning rate of $1e^{-3}$ is used for all parameters being fine-tuned with an ADAM optimiser. We train with a fixed quantity of 10 epochs, and for each epoch, we perform a fixed quantity of 50 gradient updates.

3 Experiments

Tasks: We extend the four axes of task complexity examined by Esmaeili et al. [5] to the following (conducting evaluations in the native image spaces): **(i)** Variation in image voxel count (i.e. volume size), **(ii)** image spacing/anisotropy, **(iii)** target geometry (spherical vs irregular vs jagged targets occupying low-volume fractions), **(iv)** target size **(v)** target detectability and **(vi)** training dataset size. In line with Esmaeili et al. [5] we chose binary semantic segmentation tasks from the Medical Segmentation Decathlon (MSD) [2]. For multi-sequence datasets, the sequence that best visualises the target was used.

Table 1. Task characteristics. All tasks are binary semantic segmentation from the MSD [2]. Dataset sizes are post-split.

Task	Sequence	Key characteristics	N
Hippocampus	T1	Small volume, fine detail	130
Brain tumour core	T2w	Irregular, ambiguous boundaries	240
Pancreas	CT	Large, blobby	140
Liver	CT	Very large, blobby, easy detection	65
Prostate	T2w	Spherical, highly anisotropic	16
Lung lesion	CT	Small target in large volume	31
Hepatic vessels	CT	Sparse, branching, low volume fraction	151
Colon cancer	CT	Hard to detect	63

Prompting: As with Esmaeili et al. [5], we simulate one point per class for each interaction step. We sample points with uniform random sampling from

the reference annotation for initialisation, and the the false-negative error regions computed between predictions and a reference annotation for edits. Simulations consist of interactive initialisation and 100 editing steps.

Model Evaluation: We extend the evaluation metrics of Esmaeili et al. [5]. On a per-sample basis, we compute Dice and Normalised Surface Dice (NSD) with MSD [2] tolerances at each interaction step. Using these we obtain AUCs normalised by interaction count for both metrics (nAUC), this metric captures editing stability and speed of convergence to the performance ceiling. Following Esmaeili et al. [5], we estimate the number of interactions to expert-level performance (NoI) on a per-sample Dice basis, normalised by the maximum interaction count to obtain a percentage (nNoI). A proxy for expert-level performance is defined as the task-specific mean Dice of nnU-Net [6] trained on the full training set with optimal configuration selection, as nnU-Net is a state-of-the-art segmentation algorithm used in many annotation pipelines. All expected metric values are reported as dataset-wide means. The percentage of samples that did not reach the performance target (NoF) is also reported. For adaptation studies, we extend the evaluation by tracking the trajectory of each metric's expected performance $\mathbb{E}[m_j(t)]$ as a function of dataset size t, evaluating after each training episode. We report AUCs over these trajectories (trajectory AUCs) to get a measure of the overall performance trajectory as samples are provided in the data stream. Significance rankings follow the MSD protocol [2] with pairwise algorithm comparisons: Wilcoxon signed-rank tests for all metrics, except episodic NoF where a McNemar test is used since comparisons are between binary outcomes. Significance is reported at the $\alpha = 0.05$ level. We elect not to perform significance tests for trajectory measures as sample pairs constructed from data-sample index are not independent due to their construction. Moreover, with a single split the power of significance tests computed on AUC metrics would be low.

Data Splitting and Evaluation Runs: MSD training datasets were split 50-50 into a train and holdout set. For training runs the data stream is drip-fed to form the annotation cache in Section.2; with the sequence of the training data stream permuted across 3 runs. Using this, we obtain 3 runs with episodic checkpoints. Inference is performed on 3 runs to simulate prompting stochasticity using the holdout set. For adaptive methods we perform hold-out inference runs on every model checkpoint in the corresponding training run. For static models we use the same pre-trained checkpoint across inference runs. For episodic performance measures we first average metrics across runs before reporting statistics. For trajectory measures we map the expected performance at discrete training episodes onto the corresponding interval of the training process, producing a continuous curve that describes performance over time. We average trajectories across training runs, as training is not always triggered synchronously across runs.

4 Results and Discussion

Table 2 compares nnInteractive and CLoPA-I.N after the final training episode (approximately nnU-Net-equivalent quantity of training data).

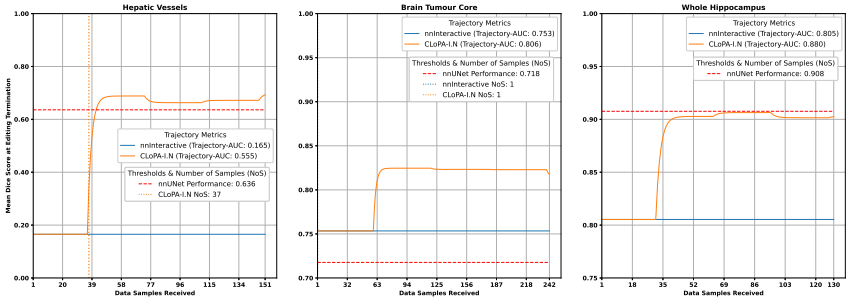
Table 2. Final-episode performance. CLoPA-I.N: instance normalisation (I.N) only. All metrics report means except NoF (percentage of failed samples). Bold indicates first-ranked method per metric and task according to significance rankings.

Task	Algorithm	Dice			NSD			NoI	
		Init.	Iter. 100	nAUC	Init.	Iter. 100	nAUC	nNoI	NoF
Brain Tumour	nnInteractive	0.505	0.753	0.742	0.583	0.922	0.904	29.6	24.4
	CLoPA-I.N	**0.682**	**0.815**	**0.817**	**0.812**	**0.968**	**0.963**	**14.8**	**12.4**
Liver	nnInteractive	0.373	0.970	0.963	0.384	0.992	0.984	22	18.2
	CLoPA-I.N	**0.912**	**0.974**	**0.973**	**0.910**	**0.994**	**0.992**	**11.7**	**6.06**
Hippocampus	nnInteractive	0.585	0.805	0.788	0.631	0.890	0.870	99	97.7
	CLoPA-I.N	**0.875**	**0.903**	**0.902**	**0.957**	**0.984**	**0.983**	**45.6**	**40**
Prostate	nnInteractive	0.774	0.922	0.921	0.815	0.985	0.983	4.6	0
	CLoPA-I.N	**0.882**	**0.935**	**0.932**	**0.941**	**0.992**	**0.990**	**1.8**	0
Lung Lesion	nnInteractive	0.698	0.849	0.852	0.743	0.940	0.936	2	0
	CLoPA-I.N	**0.764**	**0.861**	0.856	**0.819**	0.950	0.942	1.5	0
Pancreas	nnInteractive	0.454	0.895	0.881	0.538	0.983	0.971	9.8	2.13
	CLoPA-I.N	**0.668**	**0.897**	**0.888**	**0.760**	**0.985**	**0.979**	**8.3**	2.84
Hepatic Vessels	nnInteractive	0.113	0.165	0.209	0.122	0.232	0.289	84.7	82.9
	CLoPA-I.N	**0.511**	**0.698**	**0.687**	**0.682**	**0.871**	**0.856**	**19.6**	**11.8**
Colon	nnInteractive	0.475	0.728	0.725	0.520	0.788	0.790	2.4	0
	CLoPA-I.N	**0.626**	**0.819**	**0.813**	**0.721**	**0.913**	**0.911**	**1.2**	0

Tasks where the base model converges: For tasks where nnInteractive already achieves consistent convergence to nnU-Net-level performance (NoF $\lesssim 5\%$: liver, prostate, pancreas, lung lesion, colon cancer)—typically blobby targets or those where detectability is the main challenge—CLoPA-I.N maintains convergence rates (except pancreas, though not at a statistically significant rate) while improving all other metrics. Initialisation Dice and NSD are substantially higher, indicating that task-alignment accelerates annotation efficiency, especially for large targets like liver where adaptation provides sufficient context to trigger nnInteractive's auto-zoom mechanism with minimal prompting. Final-iteration Dice and NSD, as well as nAUC metrics, also improve—particularly for tasks where zero-shot performance was not already saturated, such as colon cancer. Thus, task-alignment raises the performance ceiling and improves editing stability even for easier tasks. For these tasks CLoPA-I.N is likely effective because the existing features are sufficient for blobby targets in moderately sized images. With relatively fewer samples (all these tasks are in medium dataset size at most) instance normalisation tuning is also robust against overfitting.

Tasks where the base model struggles: For more challenging tasks (NoF $\gtrsim$ 20%: brain tumour core, hippocampus, hepatic vessels), adaptation yields large performance gains across all metrics: substantially better initialisation, faster annotation (nNoI), more robust editing stability (nAUC), and higher performance ceilings. These tasks share characteristics that expose limitations in nnInteractive's zero-shot capabilities: ambiguous segmentation boundaries (brain tumour core), fine detail requirements at image sizes deviating from the model's 192^3 input patch (brain tumour core, hippocampus), and sparse branching structures

Fig. 2. Comparison across methods for the trajectory of mean Dice Score after the editing terminates, as a function of data samples received. Left to right, trajectory on tasks: hepatic vessel, brain tumour core, hippocampus. Vertical lines indicate the number of samples (NoS) to reach nnU-Net performance (horizontal red line).



with low volume fractions (hepatic vessels). For brain tumour core and hippocampus, the base model plateaus early but remains stable (nAUC slightly below final-iteration performance), indicating a performance ceiling when segmenting ambiguous targets with sparse point prompts. Instance normalisation adaptation partially mitigates this but improvements plateau after the first training episode (Fig. 2). This is likely because the feature representations were not adjusted for ambiguous targets or images smaller than the patch size.

Hepatic vessels present a different challenge: sparse branching structures with low volume fractions, likely underrepresented in nnInteractive's training data. The base model exhibits volatile peak-and-dip editing behaviour (nAUC much higher than final-iteration metrics). Post adaptation, initialisation performance immediately spikes and the editing performance does not exhibit this behaviour. After the first training episode (Fig. 2), instance normalisation tuning allows expected post-editing performance to exceed nnU-Net performance within approximately 20% of the editing budget. However, with NoF at 11.8%, performance still saturates—as illustrated by the plateauing after the initial episode in Fig. 2. This saturation likely reflects the inability of instance normalisation alone to learn task-specific feature representations. Contributing factors may include the CNN architecture's limitations for targets with long-range dependencies (suggesting a necessity for tuning the convolution kernels).

Trajectory analysis: Table 3 confirms the same trends across the full dataset-size trajectory, though improvements are less pronounced than in the final-episode snapshot because the trajectory AUC averages over all episodes, including before adaptation takes effect. As shown in Fig. 2, a crude method which only tuned a tiny fraction of the base model's parameters ($<0.01\%$) was capable of exceeding, or coming extremely close nnU-Net performance on unattainable tasks for the base model (hepatic vessel and hippocampus). Moreover, we see that instance normalisation tuning rapidly elevates expected performance to near expert-level after the first training episode, but subsequently plateaus across all tasks. This is practically significant: it means that clinicians benefit from im-

Table 3. Trajectory AUC performance across all tasks. CLoPA-I.N: instance normalisation (I.N) only. Metrics represent the AUC computed on the trajectory of expected performance for each metric as a function of dataset size.

Task	Algorithm	Dice			NSD			NoI	
		Init.	Iter. 100	nAUC	Init.	Iter. 100	nAUC	nNoI	NoF
Brain Tumour	nnInteractive	0.505	0.753	0.742	0.583	0.922	0.904	26.8	20.4
	CLoPA-I.N	0.623	0.806	0.802	0.740	0.960	0.950	15.8	11.7
Liver	nnInteractive	0.373	0.970	0.963	0.384	0.992	0.984	17.5	13.6
	CLoPA-I.N	0.734	0.972	0.968	0.737	0.992	0.988	15.8	10.2
Hippocampus	nnInteractive	0.585	0.805	0.788	0.631	0.890	0.870	98	94.9
	CLoPA-I.N	0.803	0.880	0.875	0.874	0.962	0.956	55.4	48.6
Prostate	nnInteractive	0.774	0.922	0.921	0.815	0.985	0.983	4.2	0
	CLoPA-I.N	0.850	0.932	0.929	0.899	0.991	0.988	3.7	0.694
Lung Lesion	nnInteractive	0.698	0.849	0.852	0.743	0.940	0.936	1.8	0
	CLoPA-I.N	0.748	0.860	0.857	0.797	0.947	0.939	1.7	0
Pancreas	nnInteractive	0.454	0.895	0.881	0.538	0.983	0.971	9.3	2.13
	CLoPA-I.N	0.593	0.899	0.887	0.683	0.986	0.977	8.1	2.01
Hepatic Vessels	nnInteractive	0.113	0.165	0.209	0.122	0.232	0.289	82	79.8
	CLoPA-I.N	0.407	0.555	0.556	0.542	0.716	0.715	39.4	31.7
Colon	nnInteractive	0.475	0.728	0.725	0.520	0.788	0.790	2.4	0
	CLoPA-I.N	0.576	0.783	0.779	0.660	0.867	0.867	1.7	0

proved predictions early in the annotation campaign, reducing cumulative user effort over the full dataset. Since the initial episode yields the majority of the benefit, a two-phase strategy—triggering instance normalisation adaptation early, then transitioning to feature-representation tuning as more data becomes available—would likely yield further gains. However, configuring such a curriculum would likely depend on task characteristics and the size of the cached annotated data. This reflects both the magnitude of representational change required relative to the current solution, and the degree to which additional annotations constrain the space of plausible updates. Such a curriculum could also incorporate dynamic prompting strategies, for instance extending the interaction window, to maximise the quality of the training signal for longer editing sequences. In summary, low-parameter adaptation elevates performance ceilings, stabilises editing behaviour, boosts annotation efficiency and enables specialist performance to be reached using low-effort clicking, with only a fraction of the data.

An open question concerns what CLoPA is actually learning. In annotation, the goal is to recover underlying anatomy, not the idiosyncratic decisions of any individual annotator. Since MSD only contains consensus-fused labels, the current evaluation cannot distinguish between these two cases: improvement on held-out samples may reflect true anatomical learning, or simply fitting to one annotation style. A multi-annotator study would resolve this directly, by assessing whether adaptation on per-annotator perspectives converges toward inter-annotator consensus — the cleaner proxy for ground truth.

Acknowledgments. Parhom Esmaeili is supported by the Engineering and Physical Sciences Research Council iCASE (grant number EP/Y528572/1) and co-funded by Siemens Healthineers.

Disclosure of Interests. The authors declare no competing interests to report.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (11 2012). https://doi.org/10.1109/TPAMI.2012.120
2. Antonelli, M., Reinke, A., Bakas, S., et al.: The Medical Segmentation Decathlon. Nature Communications **13**(1), 4128 (7 2022). https://doi.org/10.1038/s41467-022-30695-9
3. Cheng, J., Ye, J., Deng, Z., Chen, J., Li, T., Wang, H., Su, Y., Huang, Z., Chen, J., Jiang, L., Sun, H., He, J., Zhang, S., Zhu, M., Qiao, Y.: SAM-Med2D (8 2023)
4. Du, Y., Bai, F., Huang, T., Zhao, B.: SegVol: Universal and Interactive Volumetric Medical Image Segmentation. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. pp. 110746–110783. Curran Associates Inc, Vancouver (2024)
5. Esmaeili, P., Borges, P., Fernandez, V., Gibson, E., Ourselin, S., Cardoso, M.J.: A Methodology for Clinically Driven Interactive Segmentation Evaluation. In: Guo, X., Jin, Y., Lamdouar, H., Men, Q., Ouyang, C., Sahu, M., Vedula, S.S. (eds.) International Workshop on Human-AI Collaboration. pp. 13–22. Springer Nature Switzerland, Cham (2026). https://doi.org/10.1007/978-3-032-08970-0{_}2
6. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2 2021). https://doi.org/10.1038/s41592-020-01008-z
7. Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., Hamm, B., Wald, T., Langenberg, M., Ulrich, C., Deissler, J., Floca, R., Maier-Hein, K.: nnInteractive: Redefining 3D Promptable Segmentation. arXiv preprint (3 2025)
8. Jadon, S.: A survey of loss functions for semantic segmentation. In: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB). pp. 1–7. IEEE, Via del Mar (10 2020). https://doi.org/10.1109/CIBCB48159.2020.9277638
9. 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 preprint (4 2023)
10. de Kok, J.W.T.M., de la Hoz, M.A.A., de Jong, Y., et al.: A guide to sharing open healthcare data under the General Data Protection Regulation. Scientific Data **10**(1), 404 (6 2023). https://doi.org/10.1038/s41597-023-02256-2
11. Lenchik, L., Heacock, L., Weaver, A.A., et al.: Automated Segmentation of Tissues Using CT and MRI: A Systematic Review. Academic Radiology **26**(12), 1695–1706 (12 2019). https://doi.org/10.1016/j.acra.2019.07.006
12. Ravi, N., Gabeur, V., Hu, Y.T., et al.: SAM 2: Segment Anything in Images and Videos. arXiv preprint (8 2024)

13. Ulyanov, D., Vedaldi, A., Lempitsky, V.: Instance Normalization: The Missing Ingredient for Fast Stylization. arXiv preprint (11 2017)
14. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: SAM-Med3D. arXiv preprint (10 2023)
15. Wong, H.E., Ortiz, J.J.G., Guttag, J., Dalca, A.V.: MultiverSeg: Scalable Interactive Segmentation of Biomedical Imaging Datasets with In-Context Guidance. arXiv preprint (12 2024)
16. Wong, H.E., Rakic, M., Guttag, J., Dalca Adrian V.: ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image. In: European Conference on Computer Vision (ECCV). Milan (2024)
17. Xu, W., Liang, Z., Anthony, H., Ibrahim, Y., Cohen, F., Yang, G., Kamnitsas, K.: You Point, I Learn: Online Adaptation of Interactive Segmentation Models for Handling Distribution Shifts in Medical Imaging. In: The Fourteenth International Conference on Learning Representations (12 2025)