IMVS: Interactive Medical Volume Segmentation with Test-Time Adaptation - A New Method for Annotating Radiology Datasets

Abhilaksh Singh Reen[⋆,1(✉)], Kushal Borkar[⋆,2], and Ritvik Mahapatra[⋆,3]

[1] Independent Researcher, Delhi, India
abhilakshsinghreen@gmail.com
[2] Independent Researcher, Kolkata, India
kushalborkar.11@gmail.com
[3] California State University, Fresno, CA, USA
ritvik123@mail.fresnostate.edu

Abstract. Annotating large radiology datasets is bottlenecked by the manual effort of delineating structures slice-by-slice in 3D volumes. Interactive methods reduce this effort but stay interaction-inefficient: slice-wise methods (including many foundation models) ignore inter-slice continuity, while 3D and video-based methods propagate a prompt with a *fixed* propagator that never adapts to the target volume, so it drifts on low-contrast or pathological structures and must be re-prompted. We present IMVS, a human-in-the-loop annotation framework that composes three components into a closed loop rather than a new segmentation primitive: a lightweight 2D Slice Mask Adapter (SMA) fine-tuned online from user scribbles, a frozen Volume Mask Tracker (VMT) that propagates corrected masks across adjacent slices, and a soft teacher–student alignment that limits forgetting. The SMA is backbone-agnostic (UNet++, DeepLabV3, TransUNet). Across 8 public CT/MRI datasets, IMVS matches strong interactive baselines in quality while sharply cutting annotation effort: 14.4× faster than a proficient copy-based manual workflow (22.3× over naive manual), 4.6× over slice-wise and 1.9× over 3D interactive methods. MedSAM2 and ScribblePrompt stay competitive or stronger on well-delineated organs; IMVS's advantage is largest on challenging targets and on interaction efficiency. Source code and Demo Video: https://github.com/AbhilakshSinghReen/imvs.

Keywords: Human-in-the-Loop Annotation · Interactive Segmentation · Annotation Efficiency · Test-Time Adaptation · Medical Image Analysis

1 Introduction

Interactive segmentation frameworks, including foundation models such as SAM [9], MedSAM [15], and ScribblePrompt [24], cut annotation effort by taking human prompts (boxes [26], clicks [19, 20, 14], or scribbles [25, 1]) into the inference

⋆ Equal contribution.

loop. Applied to volumes, slice-wise methods refine masks well but ignore inter-slice continuity, producing redundant interactions. Recent 3D and video-based methods—PRISM [11], nnInteractive [6], and video-propagation systems such as SAM2/MedSAM2 [16]—carry a prompt through the volume but rely on a *fixed* propagator never adapted to the volume being annotated. Under the distribution shift, low contrast, and pathology typical of clinical targets, they drift and must be re-prompted, so effort goes into correcting the same errors repeatedly.

Motivated by the temporal consistency exploited in Video Object Segmentation (VOS), our framework IMVS likewise propagates masks across adjacent slices, but unlike the fixed 3D propagators above, couples propagation with *online adaptation*: each correction fixes the current slice *and* updates a lightweight per-slice model for the slices ahead, so an interaction is reused over many slices instead of repeated on each one. In the taxonomy of interactive medical segmentation [17], IMVS is scribble-driven and iteratively-refined, distinguished by *where* human effort is spent: it amortizes corrections over a volume through learned propagation and online adaptation. We position it not as a new segmentation primitive but as a *human–AI collaboration system* for annotation efficiency: its building blocks—interactive 2D segmentation, teacher–student test-time adaptation [21], soft alignment against forgetting [10], and attention-based propagation—are established; the contribution is how they compose into an efficient closed loop.

We contribute: (1) a closed-loop pipeline coupling a lightweight, online-adapted 2D Slice Mask Adapter (SMA) with a frozen Volume Mask Tracker (VMT), confining backpropagation to a small network while a fixed propagator enforces consistency; (2) a soft teacher–student alignment enabling refinement from user scribbles without forgetting across slices and volumes; and (3) an empirical study on 8 CT/MRI datasets with 9 annotators, quantifying annotation-efficiency gains and characterizing where the approach helps and where it does not. The SMA is backbone-agnostic (UNet++, DeepLabV3, TransUNet).

2 Related Work

Interactive methods refine masks from user inputs: boxes [26], clicks [19, 20, 14], or scribbles [25, 1]; f-BRS [19] and RiTM [20] introduced iterative inference-time refinement, and iSegFormer [14] a transformer backbone for medical images. Foundation models (SAM [9], MedSAM [15, 16]) generalize across targets but degrade under medical distribution shift and are typically applied per-slice; lighter 3D approaches such as PRISM [11] and promptable systems like nnInteractive [6] address volumes directly. In Marinov et al.'s taxonomy [17], IMVS is scribble-driven and iteratively-refined, distinguished by amortizing interactions across slices rather than maximizing single-mask accuracy [24, 13].

Test-time and continual adaptation avoid target-domain retraining using uncertainty [21, 12] or teacher pseudo-labels [23]; mean-teacher formulations are effective for continual TTA [4] but are *unsupervised*. IMVS is instead *interactive and weakly supervised*: the dominant signal is the user's corrective scribble, with

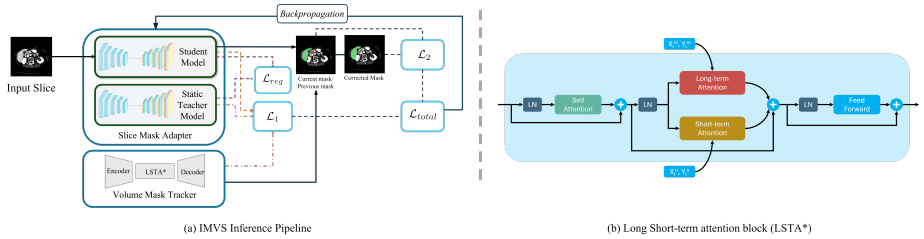

Fig. 1. The IMVS annotation loop: the SMA predicts a mask the user refines with scribbles; the frozen VMT propagates the correction across adjacent slices.

teacher/propagation terms only stabilizing, countering the pseudo-label drift to which unsupervised teacher–student adaptation is prone [23], while soft alignment guards against forgetting [10, 5]. Propagating an annotation to neighbouring slices is long-standing—Wang et al. [22] used online random forests for placental segmentation, and VOS trackers later formalized memory/attention-based propagation. Like modern VOS trackers, IMVS uses a learned attention-based propagator with a memory bank, pretrained on video object segmentation [27] and fine-tuned for medical propagation. What distinguishes IMVS is that this propagator is kept lightweight and frozen during the loop, its errors recovered by online SMA adaptation rather than a heavier tracker. This division of labour—a fixed propagator plus an adaptive per-slice model—is the main design choice.

3 Method

Problem setting and loop. IMVS couples a lightweight 2D SMA, fine-tuned online from scribbles, with a frozen VMT propagator, confining backpropagation to the small SMA while the fixed propagator enforces volumetric consistency. Let f_{θ_0} be a model pre-trained (parameters θ_0) on source data D_s. At step t the model sees a slice $x_t \in \mathbb{R}^{H\times W}$, the previous slice/mask $(x_{t-1}, \hat{y}_{t-1})$, and (on intervention) scribbles $\mathcal{S}_t$, and the user refines the prediction into an accepted mask m_t; the goal is to segment an unseen volume with as few interactions as possible without forgetting D_s. The annotator begins at the slice x_0 where the target first becomes clearly visible - a *protocol* choice mirroring how experts scroll to a structure's onset, not a tuned hyperparameter. Because the loop re-seeds from *any* corrected slice, a different start costs at most one early correction, not final quality. As shown in Fig. 1, from x_0 the framework sweeps the volume forward only, alternating an *Interactive Mode* (scribbles drive online SMA adaptation) and an *Automatic Mode* (the VMT propagates the accepted mask forward for up to K slices before a correction is required). We use K=17, trading propagation error against correction frequency; K is justified empirically in Sec. 4.4. Algorithm 1 summarizes the loop: the user intervenes only when a propagated mask is unsatisfactory, and each correction fixes the current slice and adapts the student for the slices ahead.

3.1 Slice Mask Adapter (SMA)

The SMA is a standard, interchangeable 2D segmentation network; we evaluate UNet++ [30], DeepLabV3+, and TransUNet, and adopt UNet++ as default. It is instantiated as a teacher–student pair: a frozen teacher f_{θ^T} and an adaptive student f_{θ^S}, both initialized from the same weights $\theta^T = \theta^S = \theta_0$. The teacher preserves source-domain knowledge; only the student is updated online. The teacher and the VMT are *never* updated at test time. The user's accepted refinement of the student's prediction is $\hat{y}_t^{corrected}$; each update improves that prediction at the *next* correction, not the current slice.

How scribbles enter (early fusion). Foreground/background scribbles $\mathcal{S}_t = \{s_t^{fg}, s_t^{bg}\}$ are rasterized (dilation radius 3 px) into two binary guidance channels $M_{fg}, M_{bg} \in \{0,1\}^{H\times W}$, set to 1 on foreground/background strokes respectively, and *concatenated with the grayscale slice at the input*, giving a 3-channel input (CT/MRI + $M_{fg} + M_{bg}$) to the SMA. Interactions are thus fused early, keeping them backbone-agnostic.

3.2 Volume Mask Tracker (VMT)

The VMT maps $(x_t, m_{t-1}, \mathcal{M}) \mapsto \hat{y}_t^{VMT}$, propagating the accepted mask of slice $t-1$ onto slice t using a memory bank $\mathcal{M}$ of recent (feature, mask) pairs. Its encoder is a **ViT-B/16** pretrained on video object segmentation (YouTube-VOS 2019 [27]) and fine-tuned on medical volumes (Sec. 3.5), producing $F_t^{enc} = \text{Encoder}(x_t) \in \mathbb{R}^{H/16\times W/16\times D}$ with $D = 768$. The previous mask is encoded to the same resolution (a 1×1 convolution on the resized mask) and concatenated with slice features to inject object-specific context. A transformer with $L = 6$ Long+Short-Term Attention (LSTA) [28, 29] layers applies scaled dot-product attention in two scopes: *short-term* over the two preceding slices (smoothness) and *long-term* over a memory buffer of the six most recent slices (appearance consistency and drift prevention). A decoder upsamples $\hat{y}_t^{VMT}$. VMT is frozen at inference.

3.3 Online Adaptation Objective

We call this *test-time adaptation* to indicate *when* the SMA is updated (at inference, on the target volume), not unsupervised adaptation [21, 4]: the dominant signal is the user's corrective scribble, with propagated/teacher pseudo-labels only for stabilization. At each interactive step the student minimizes three terms. (A) A *consistency* term distills the frozen VMT mask and the augmentation-averaged teacher prediction $\hat{y}_t^{aug}$ into the student $\hat{y}_t^S$,

$$\mathcal{L}_1 = -\sum_c \big[\, \beta_1\, \hat{y}_t^{VMT}[:,c] \log(\hat{y}_t^S[:,c]) + \beta_2\, \hat{y}_t^{aug}[:,c] \log(\hat{y}_t^S[:,c])\big], \qquad (1)$$

with class index c and weights β_1, β_2 biased toward VMT guidance. (B) An *interactive* cross-entropy term $\mathcal{L}_2 = -\sum_c \hat{y}_t^{corrected}[:,c] \log(\hat{y}_t^S[:,c])$ matches the

Algorithm 1 IMVS interaction loop for one volume

Require: volume $\{x_i\}$; SMA f_{θ_0}; frozen VMT g; threshold γ; span K
$\theta^T, \theta^S \leftarrow \theta_0$; user seeds x_0 with $\mathcal{S}_0$ (Interactive Mode); $\hat{y}_0^S \leftarrow f_{\theta^S}(x_0, \mathcal{S}_0)$; refine and update θ^S; $m_0 \leftarrow \hat{y}_0^{corrected}$; push to $\mathcal{M}$
for $t = 1, 2, \ldots$ until the volume is covered **do**
  $\hat{y}_t^{VMT} \leftarrow g(x_t, m_{t-1}, \mathcal{M})$ ▷ automatic propagation
  if user accepts $\hat{y}_t^{VMT}$ **and** span since correction $< K$ **then**
    $m_t \leftarrow \hat{y}_t^{VMT}$
  else ▷ correct and adapt
    user draws $\mathcal{S}_t$; $\hat{y}_t^S \leftarrow f_{\theta^S}(x_t, \mathcal{S}_t)$; if $\mathcal{L}_{total} > \gamma$ update θ^S; $m_t \leftarrow \hat{y}_t^{corrected}$
  end if
  push (F_t, m_t) to $\mathcal{M}$, evict oldest
end for
carry θ^S to the next volume ▷ warm-start

user-corrected mask $\hat{y}_t^{corrected}$, and (C) a soft *alignment* term prevents forgetting by anchoring the student's batch-norm (BN) parameters to the teacher, $\mathcal{L}_{reg}(\theta_t) = \sum_l \mathbb{1}[l \in \text{BN}] \, \|\theta_l^S - \theta_l^T\|_2^2$. The total objective is

$$\mathcal{L}_{total} = \mathcal{L}_1 + \mathcal{L}_2 + \kappa \, \mathcal{L}_{reg}. \tag{2}$$

The student is updated by SGD only when $\mathcal{L}_{total} > \gamma = 0.05$ (suppressing noisy updates), with $\kappa = 0.1$. Freezing the propagator prevents pseudo-label drift from compounding; the adapted student is carried to subsequent volumes as an efficiency warm-start (supervision stays user-driven, no bias to annotation).

3.4 Human-AI Interaction Workflow

IMVS is built around iterative human-AI collaboration rather than one-shot segmentation. The annotator intervenes only when propagated masks become unacceptable; each correction is immediately incorporated by the adaptive SMA and reused across subsequent slices through propagation, a capability missing in existing methods. User input thus serves not only as local error correction but as guidance that adapts future model behavior, reducing repeated corrections within a session, and shifts the annotator's role from refining every slice to supervising an adaptive segmentation assistant.

3.5 Training and Implementation

Data / zero-shot split. Both networks are trained *only* on stratified 70:15:15 splits of three datasets (BraTS [18], LiTS [3], MSD Pancreas [2]; the "consolidated" set); the other five (CHAOS-CT/MRI [8], AMOS-CT [7], MSD Prostate, MSD HepaticVessel) are never seen in training and evaluated **zero-shot**, so cross-dataset results measure generalization. **SMA:** ImageNet-initialized backbones, slices 256×256, 110 epochs (AdamW, lr $1{\times}10^{-4}$, batch 16), loss $\mathcal{L}_{Dice} +$

Table 1. Annotation efficiency per method and dataset: **time** (minutes) with # **interactions** in parentheses. Manual baselines have no discrete interactions (−). Best per column in bold.

Method	CHAOS CT	CHAOS MRI	AMOS CT	MSD Prostate	LiTS	MSD Hep.Ves.	MSD Pancreas	BraTS
Manual	44.79 (−)	21.5 (−)	65.5 (−)	8.53 (−)	73.53 (−)	96.3 (−)	8.11 (−)	55.26 (−)
Manual w/ Copy	28.83 (−)	13.68 (−)	42.79 (−)	5.4 (−)	47.56 (−)	61.85 (−)	5.26 (−)	35.52 (−)
f-BRS	7.23 (10)	3.41 (9)	10.55 (34)	1.64 (16)	17.57 (27)	22.8 (126)	1.59 (23)	11.49 (30)
Med SAM	12.61 (13)	6.02 (11)	18.51 (31)	2.39 (18)	20.51 (28)	28.25 (122)	2.27 (22)	15.55 (30)
ScribblePrompt	7.79 (20)	3.65 (12)	11.36 (29)	1.45 (21)	12.68 (25)	18.4 (121)	1.39 (24)	9.41 (29)
iSegFormer	7.12 (8)	3.36 (10)	10.42 (10)	1.35 (6)	11.58 (10)	15.4 (32)	1.32 (8)	8.73 (9)
PRISM	5.49 (6)	2.65 (4)	8.06 (9)	1.07 (3)	3.93 (5)	14.7 (28)	1.03 (3)	6.79 (8)
MedSAM2	**1.89** (**3**)	**0.96** (**2**)	**2.74** (**3**)	0.93 (3)	4.17 (5)	11.3 (**19**)	0.76 (3)	3.1 (4)
nnInteractive	1.94 (4)	0.965 (3)	2.81 (4)	0.7 (3)	3.705 (**4**)	11 (23)	0.61 (4)	2.75 (5)
Ours (UNet++)	1.99 (4)	0.97 (3)	2.88 (4)	**0.47** (**2**)	**3.24** (**4**)	**10.7** (20)	**0.46** (**2**)	**2.4** (**3**)

$\mathcal{L}_{CE}$; training scribbles simulated from ground truth (boundary default). Online adaptation uses SGD (momentum 0.9, lr 1×10^{-4}) with horizontal-flip and intensity-jitter test-time augmentation on the teacher. **VMT:** the ViT-B propagator is pretrained on YouTube-VOS 2019 [27] and fine-tuned as a mask propagator on the consolidated set for 58 epochs (Adam, lr 2×10^{-4}, batch 8) with a CE+Dice loss between propagated mask and ground truth, then frozen at inference. **Hardware:** a single NVIDIA RTX 5090 (CUDA 12.8).

Training uses *simulated* scribbles: a deterministic simulator reacts to the current error region until the stopping criterion is met, which is reproducible and annotator-independent.

4 Experiments and Results

We evaluate across **8 CT/MRI benchmark datasets**, focusing on challenging cases (liver tumors, pancreatic neoplasms, hepatic vasculature); training uses only the BraTS/LiTS/MSD-Pancreas splits, the other five being held out zero-shot. We benchmark seven interactive methods: MedSAM, MedSAM2, ScribblePrompt, PRISM, f-BRS, iSegFormer, and nnInteractive [6]. Interaction modality is heterogeneous (MedSAM/MedSAM2 and PRISM use boxes/clicks; ScribblePrompt, f-BRS, iSegFormer, and IMVS use scribbles/clicks; nnInteractive supports all three), which we account for when attributing efficiency gains.

4.1 User Study and Stopping Criterion

Annotation times were measured in a user study with **nine residents** of varying experience, a realistic annotator population for radiology dataset construction under expert supervision. **Protocol.** Annotators worked in one of two interfaces—an in-house 3D Slicer extension or a cloud-based viewer built on OHIF—under a single acceptance rule: advance once the segmentation is qualitatively satisfactory. All nine annotated the same volumes under every method, and reported times are the mean over the nine. Aggregated accepted masks defined the automatic stopping threshold (DSC = 0.885, NSD = 0.894, HD95 = 4 mm). Inter-annotator agreement was high (mean pairwise DSC = 0.89 ± 0.04), with consistent thresholds across raters (DSC = 0.885 ± 0.03), supporting its use as a shared comparison rule. The study (Table 1) shows workload reduction in both time and interactions.

Table 2. Interaction efficiency at matched quality: interactions needed to reach DSC ≥ 0.90 per dataset; lower is better. Complements Table 1 (effort to satisfy the stopping criterion). IMVS needs the fewest on every dataset.

Method	CHAOS-CT	CHAOS-MRI	AMOS	Prostate	LiTS	Pancreas	BraTS	Mean
iSegFormer	13	14	13	12	11	15	13.5	13.1
PRISM	12	13	12	11	10	14	12.5	12.1
MedSAM2	10	11	10	9	8	12	10.5	10.1
Ours	**7**	**8**	**7**	**6**	**6**	**8**	**7.5**	**7.1**

4.2 Annotation Efficiency

Time and interactions. Table 1 reports annotation time and interaction counts. IMVS is fastest on 5 of 8 datasets and in aggregate, but not uniformly best. nnInteractive is the closest competitor; it and MedSAM2 are faster on the well-delineated organs of CHAOS-CT/MRI and AMOS-CT, by at most 0.07 min, reported per-dataset rather than averaged away. On interaction count IMVS is most efficient on 4 of 8 datasets (2 on MSD Prostate/Pancreas, 3 on BraTS, 4 on LiTS), never exceeds nnInteractive on any, and is within one interaction of MedSAM2 elsewhere. Tortuous hepatic vessels are the hard case for *every* method: IMVS still needs 20 interactions (vs. 19 for MedSAM2 and 28–126 for the slice-wise baselines), consistent with our claim that IMVS helps least on discontinuous structures.

Sources of the gain. We state speedups conservatively: IMVS is 22.3× faster than naive **Manual** annotation but **14.4**× faster than the fairer **Manual With Copy** (copy-and-edit the previous mask). Part of the per-interaction gain over click/box baselines is modality (a scribble carries more information than a click); comparing against *scribble* baselines under the identical criterion isolates the algorithmic contribution: amortizing interactions over a propagated window plus online SMA adaptation. IMVS also uses the least compute (3.23 GB VRAM, 15.93 s GPU/volume). A paired Wilcoxon signed-rank test (per volume, MedSAM2) locates the advantage: significant over the four challenging datasets (LiTS, MSD Pancreas, MSD HepaticVessel, BraTS; $p = 0.02$) but similar over the well-delineated ones (CHAOS-CT/MRI, AMOS, MSD Prostate; $p = 0.26$).

4.3 Segmentation Quality

Table 2 complements Table 1: interactions needed to *reach* a demanding target (DSC ≥ 0.90, near the operating point). IMVS needs the fewest on every dataset: 7.1 on average vs. 10.1 for MedSAM2 and 13.1 for slice-wise iSegFormer (1.4–1.8× fewer). Final quality is closely clustered once enough interactions are spent; IMVS's contribution is reaching it *sooner*, not exceeding the plateau.

Robustness: On tortuous hepatic vessels the VMT tracks reliably for only ∼4–6 slices (vs. up to 17 on well-defined organs), so more corrections are needed and every method degrades. IMVS stays competitive because online SMA adaptation recovers tracking with typically one scribble.

Table 3. Ablation of adaptation strategies on the consolidated set (per-volume averages). *SMA (adaptive), VMT frozen* is our method; *# Tracked* is the average span before a correction.

Approach	# Int.	# Tracked	Dice	NSD	HD95	Max VRAM	GPU Time
No SMA, VMT frozen	18	3	0.86	0.87	2.7	1.7 GB	4ms
Ours (Memory Size 6)	4	12	0.92	0.91	1.6	3.23 GB	129ms
SMA (adaptive), VMT adapted	8	7	0.92	0.91	1.6	20.7 GB	1857ms

4.4 Ablation: Contribution of Each Component

Table 3 isolates the components on the consolidated set. **Online SMA adaptation** is the largest factor: without it (*No SMA, VMT frozen*) the loop needs 18 interactions at DSC 0.86; with it, 4 interactions at DSC 0.92 and a tracked span rising from 3 to 12 slices. **Freezing the VMT** is essential: adapting it as well gives no accuracy gain but 6× the VRAM, 14× the latency, and a shorter tracked span (temporal drift). Among backbones, UNet++ beats DeepLabV3 and TransUNet on all metrics (dense skip connections aid boundary refinement under sparse scribbles) and is the default; memory-bank size 6 is optimal (4–5 degrade tracking, 7–9 add compute without gains). The propagation length K=17 comes from a sweep $K \in \{5, \ldots, 25\}$: smaller K forces frequent corrections (7.2 interactions at K=5 vs. 4.1 at K=17), while larger K raises the propagation failure rate (29% at K=17 to 55% at K=25) as masks drift.

5 Scope and Limitations

IMVS segments a *single* target per pass; multi-label volumes use sequential single-label sessions sharing one adaptation context, and native multi-object propagation is future work. Efficiency comes from amortizing interactions over a propagated span, so IMVS helps least on tortuous or discontinuous structures.

6 Conclusion

We presented IMVS, a human-in-the-loop framework composing an online-adapted 2D SMA, a frozen VMT propagator, and soft teacher-student alignment into a closed annotation loop. Its benefit is annotation efficiency: propagating each correction across slices and adapting the SMA online cuts interactions, time, and VRAM. We view IMVS as a new method in human-AI collaboration for efficient dataset construction; future work includes topology-aware and multi-object propagation.

Acknowledgments. The authors thank Dr. Amit Kumar for conducting the annotation study. This work received no external funding. All experiments used publicly available datasets.

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

References

1. Agustsson, E., Uijlings, J.R., Ferrari, V.: Interactive full image segmentation by considering all regions jointly. In: CVPR. pp. 11622–11631 (2019). https://doi.org/10.1109/CVPR.2019.01189
2. 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). https://doi.org/10.1038/s41467-022-30695-9
3. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (LiTS). Medical Image Analysis **84**, 102680 (2023). https://doi.org/10.1016/j.media.2022.102680
4. Döbler, M., Marsden, R.A., Yang, B.: Robust mean teacher for continual and gradual test-time adaptation. In: CVPR. pp. 7704–7714 (2023). https://doi.org/10.1109/CVPR52729.2023.00744
5. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**(4), 128–135 (1999). https://doi.org/10.1016/S1364-6613(99)01294-2
6. 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 arXiv:2503.08373 (2025)
7. Ji, Y., Bai, H., Ge, C., Yang, J., 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. Advances in Neural Information Processing Systems **35**, 36722–36732 (2022). https://doi.org/10.52202/068431-2661
8. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: CHAOS challenge—combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021). https://doi.org/10.1016/j.media.2020.101950
9. 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. In: ICCV. pp. 4015–4026 (2023). https://doi.org/10.1109/ICCV51070.2023.00371
10. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017). https://doi.org/10.1073/pnas.1611835114
11. Li, H., Liu, H., Hu, D., Wang, J., Oguz, I.: PRISM: A promptable and robust interactive segmentation model with visual prompts. In: MICCAI. pp. 389–399. Springer (2024). https://doi.org/10.1007/978-3-031-72384-1_37
12. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: ICML. pp. 6028–6039. PMLR (2020)
13. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: ScribbleSup: Scribble-supervised convolutional networks for semantic segmentation. In: CVPR. pp. 3159–3167 (2016). https://doi.org/10.1109/CVPR.2016.344
14. Liu, Q., Xu, Z., Jiao, Y., Niethammer, M.: iSegFormer: Interactive segmentation via transformers with application to 3D knee MR images. In: MICCAI. pp. 464–474. Springer (2022). https://doi.org/10.1007/978-3-031-16443-9_45

15. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**(1), 654 (2024). https://doi.org/10.1038/s41467-024-44824-z
16. Ma, J., Kim, S., Li, F., Baharoon, M., Asakereh, R., Lyu, H., Wang, B.: Segment anything in medical images and videos: Benchmark and deployment. arXiv preprint arXiv:2408.03322 (2024)
17. Marinov, Z., Jäger, P.F., Egger, J., Kleesiek, J., Stiefelhagen, R.: Deep interactive segmentation of medical images: A systematic review and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 10998–11018 (2024). https://doi.org/10.1109/TPAMI.2024.3452629
18. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2015). https://doi.org/10.1109/TMI.2014.2377694
19. Sofiiuk, K., Petrov, I., Barinova, O., Konushin, A.: f-BRS: Rethinking backpropagating refinement for interactive segmentation. In: CVPR. pp. 8623–8632 (2020). https://doi.org/10.1109/CVPR42600.2020.00865
20. Sofiiuk, K., Petrov, I.A., Konushin, A.: Reviving iterative training with mask guidance for interactive segmentation. In: ICIP. pp. 3141–3145. IEEE (2022). https://doi.org/10.1109/ICIP46576.2022.9897365
21. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: ICLR (2021)
22. Wang, G., Zuluaga, M.A., Pratt, R., Aertsen, M., Doel, T., Klusmann, M., David, A.L., Deprest, J., Vercauteren, T., Ourselin, S.: Slic-Seg: A minimally interactive segmentation of the placenta from sparse and motion-corrupted fetal MRI in multiple views. Medical Image Analysis **34**, 137–147 (2016). https://doi.org/10.1016/j.media.2016.04.009
23. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: CVPR. pp. 7201–7211 (2022). https://doi.org/10.1109/CVPR52688.2022.00706
24. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: ScribblePrompt: Fast and flexible interactive segmentation for any biomedical image. In: ECCV. pp. 207–229. Springer (2024). https://doi.org/10.1007/978-3-031-73661-2_12
25. Wu, J., Zhao, Y., Zhu, J.Y., Luo, S., Tu, Z.: MILCut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In: CVPR. pp. 256–263 (2014). https://doi.org/10.1109/CVPR.2014.40
26. Xu, N., Price, B., Cohen, S., Yang, J., Huang, T.: Deep GrabCut for object selection. In: BMVC (2017). https://doi.org/10.5244/C.31.182
27. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: YouTube-VOS: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
28. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. In: Advances in Neural Information Processing Systems. vol. 34, pp. 2491–2502 (2021)
29. Yang, Z., Yang, Y.: Decoupling features in hierarchical propagation for video object segmentation. In: Advances in Neural Information Processing Systems. vol. 35, pp. 36324–36336 (2022). https://doi.org/10.52202/068431-2632
30. Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J.: UNet++: Redesigning skip connections to exploit multiscale features in image segmentation. IEEE transactions on medical imaging **39**(6), 1856–1867 (2019). https://doi.org/10.1109/TMI.2019.2959609