Exploiting Longitudinal Context in Clinician-Verified Interactive Lesion Tracking

Yannick Kirchhoff[1,2,3]*, Maximilian Rokuss[1,2,3]*, Daniel Philipp Mertens[5], David Füller[5], Benjamin Hamm[1,4], Andreas Schreyer[5], Oliver Ritter[5], and Klaus Maier-Hein[1,4,6]

[1] German Cancer Research Center (DKFZ) Heidelberg, Division of Medical Image Computing, Germany
[2] Faculty of Mathematics and Computer Science, Heidelberg University, Germany
[3] HIDSS4Health – Helmholtz Information and Data Science School for Health, Karlsruhe/Heidelberg, Germany
[4] Medical Faculty, Heidelberg University, Germany
[5] University Hospital Brandenburg an der Havel, Brandenburg Medical School Theodor Fontane, Germany
[6] Pattern Analysis and Learning Group, Department of Radiation Oncology, Heidelberg University Hospital, Germany
{yannick.kirchhoff,maximilian.rokuss}@dkfz-heidelberg.de

Abstract. Tracking tumor lesions across serial CT scans is essential for oncological response assessment. Existing automated methods face a fundamental trade-off: end-to-end trackers achieve high automation but offer no opportunity to correct silent tracking failures, while decoupled registration–segmentation pipelines permit user verification yet discard the lesion's prior appearance, limiting accuracy in ambiguous cases. In this work, we propose a *Verified Tracking* paradigm: a clinician verifies a registration-proposed prompt, which the model leverages alongside the baseline lesion appearance to resolve segmentation ambiguities. We present a unified framework combining early spatial prompt fusion with latent temporal difference weighting for longitudinally-informed segmentation. To address data scarcity, we leverage large-scale synthetic pretraining, proving essential for exploiting longitudinal context, improving performance by up to 4.5 Dice points over training from scratch. Our approach secured first place in the MICCAI autoPET IV challenge. We further curate and release *PanTrack,* a new longitudinal pancreatic cancer benchmark, to assess out-of-distribution generalization. Experiments show that our model outperforms prior work in both fully automatic and the proposed verified tracking setting offering a clinically safe middle ground between automation and control. Code, model and dataset will be released at `https://github.com/MIC-DKFZ/LongiSeg`.

Keywords: Longitudinal Tracking · Interactive Segmentation · Tumor Lesions · Synthetic Pretraining

* Contributed equally. Each co-first author may list themselves as lead on their CV.

1 Introduction

Longitudinal imaging is the cornerstone of oncological response assessment. With cancer incidence projected to rise 47% by 2040 [1] and CT examination volumes increasing steadily [14], radiologists face growing pressure to evaluate serial scans efficiently. Treatment response evaluation, typically governed by RECIST 1.1 guidelines [5], requires solving two distinct subtasks per lesion: *retrieval*, identifying the same structure in a follow-up scan, and *delineation*, measuring its volume change. Both remain predominantly manual, causing substantial reading time and inter-observer variability [7, 10, 13].

Existing automated approaches address this problem in complementary yet incomplete ways. **Point-only trackers** [20, 23] retrieve lesion centers, however, without volumetric delineation. **End-to-end trackers** [15] automate both retrieval and segmentation using a single baseline click, but operate as "black boxes": if the model tracks an incorrect structure or misses a splitting lesion, the resulting segmentation is clinically invalid with no opportunity for correction. **Decoupled registration-segmentaion pipelines** [6] would permit verification of the propagated point but suffer twofold: (1) registration errors frequently displace prompts beyond the tolerance of segmentation models trained on well-centered clicks [4, 8, 14, 21], and (2) treating follow-up scans in isolation discards the *longitudinal prior* (baseline appearance) needed to resolve ambiguous findings. Finally, **automated longitudinal models** [17, 18, 22] leverage this prior lesion appearance for ehnanced segmentation but lack promptability, preventing user interaction or correction at all.

To bridge these methodological gaps, we argue that clinical deployment requires satisfying three criteria simultaneously: (i) explicit use of the baseline lesion as a longitudinal prior, (ii) a clinician-correctable mechanism to guarantee correspondence when tracking fails, and (iii) robustness to off-center prompts caused by registration or rapid correction. Because existing methods satisfy at most two, we propose a novel *Verified Tracking* paradigm: registration proposes a follow-up location, a clinician verifies/corrects it, and the model segments using both the verified prompt and baseline context. This eliminates retrieval failures through minimal oversight, freeing the model to focus entirely on longitudinally-informed delineation. We present a unified framework for this workflow, combining early pixel-level prompt fusion [8] with latent-space temporal difference weighting [17]. Critically, we identify *synthetic longitudinal pretraining* as a decisive enabler: without sufficient multi-timepoint data, longitudinal architectures collapse to single-timepoint shortcuts, ignoring the baseline scan entirely. Finally, to address longitudinal data scarcity, we curate and publicly release *PanTrack* as a dedicated out-of-distribution benchmark. As public datasets with consistent, lesion-level instance annotations across multiple timepoints remain largely unavailable, this release fills a major gap, providing a valuable new resource for lesion tracking model development. Our key contributions are:

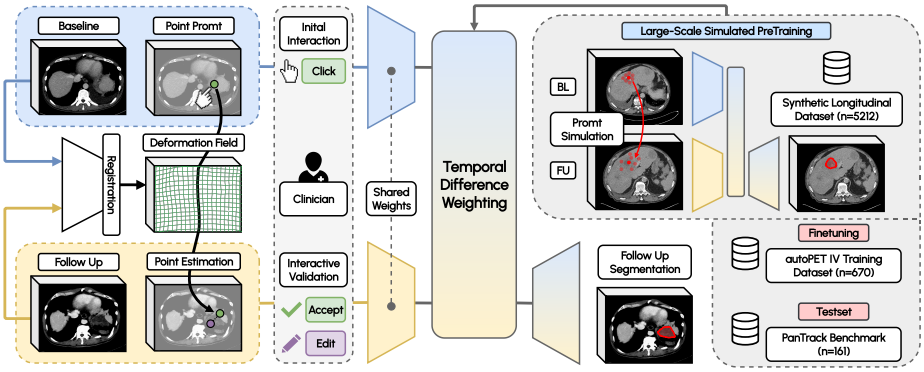

Fig. 1. Overview of our framework. The registration proposes a candidate follow-up prompt which the clinician verifies or corrects. A shared-weight encoder processes both (image, prompt) pairs and in latent space a Difference Weighting Block fuses their features by explicitly attending to temporal change before the decoder produces the longitudinally-informed follow-up segmentation. Prior, the model is pretrained on a large-scale synthetic longitudinal corpus with simulated prompts for both timepoints.

1. **Verified Tracking formulation**: We formalize a workflow where follow-up prompts are registration-proposed and optionally corrected, offering a clinically safe middle ground between automation and control.
2. **Longitudinal promptable segmentation model**: A unified architecture combining early prompt fusion and latent temporal difference weighting. Enhanced by promptable large-scale synthetic pretraining, our method won the MICCAI autoPET IV challenge, outperforming the state-of-the-art in automatic and verified tracking.
3. **The *PanTrack* benchmark**: To address multi-timepoint data scarcity, we publicly release 161 curated longitudinal CT scans (45 pancreatic cancer patients) to provide a rigorous out-of-distribution testbed for cross-domain tracking generalization and a novel model development resource.

2 Method

2.1 Problem Formulation

Given a baseline CT scan I_0 with a known lesion center p_0 and a follow-up CT scan I_t, our goal is to produce a volumetric segmentation of the corresponding lesion in I_t, conditioned on a clinician-verified follow-up point prompt p_t.

2.2 Verified Tracking via Registration

We decouple the longitudinal tracking workflow into two stages: (1) a *retrieval* step, handled jointly by a registration model and the clinician, and (2) a *delineation* step, performed by our segmentation network. For the retrieval step,

we apply uniGradICON [19], a registration foundation model trained across diverse anatomical regions, to estimate a deformation field $\phi : I_0 \to I_t$, yielding a candidate follow-up prompt:

$$\hat{p}_t = \phi(p_0). \tag{1}$$

The clinician views $\hat{p}_t$ superimposed on I_t and either accepts it or provides a corrected p_t. This verification step eliminates retrieval failures while preserving a high degree of clinical automation. For the autoPET IV dataset, the provided registration-propagated center points [12] serve directly as $\hat{p}_t$.

2.3 Segmentation Network Architecture

Unlike decoupled pipelines that segment the follow-up in isolation, our model conditions delineation on both (I_t, p_t) *and* (I_0, p_0), explicitly learning to use baseline appearance to resolve ambiguities that a single-timepoint model cannot. Our segmentation network must simultaneously handle two distinct integration challenges: incorporating a *spatial point prompt* and leveraging *baseline appearance*. These two signals empirically demand fusion at different levels of abstraction: spatial prompts are most effective at the image input level [8, 21]; longitudinal context requires latent-space integration with an explicit inductive bias towards temporal differences to prevent the network from ignoring the baseline branch altogether [3, 17]. We combine both in a Residual Encoder U-Net [9] with the hybrid fusion design below.

Early Prompt Fusion. After verified registration, we extract Volumes of Interest (VOIs) centered at p_0 and p_t from I_0 and I_t, respectively. Following prior findings that prompt encoding is most effective at the input level [8, 21], we concatenate image and prompt channel-wise:

$$X_0 = [I_0,\ G(p_0)], \quad X_t = [I_t,\ G(p_t)], \tag{2}$$

where $G(p)$ denotes a Gaussian heatmap with $\sigma = 1$ centered at p, rescaled to unit value at the center. Crucially, p_t does not need to lie precisely at the lesion center; residual registration error or clinical corrections are explicitly anticipated. Both pairs are then processed by a *shared-weight* encoder, yielding multi-scale features $\{x_0^l\}$ and $\{x_t^l\}$ at each resolution level l.

Latent Temporal Fusion via Difference Weighting. Naive channel-wise concatenation of multi-timepoint features risks the network primarily attending to a single timepoint, effectively collapsing to cross-sectional behavior [3,17]. To impose an explicit longitudinal inductive bias, we apply a Difference Weighting Block (DWB) [17] at all U-Net skip connections. For baseline and follow-up features $(x_0^l,\ x_t^l)$, the DWB computes:

$$x'^{\,l}_t = x_t^l \ \times\ \mathrm{InstNorm}\big(x_t^l - x_0^l\big) + x_t^l. \tag{3}$$

The normalized feature difference acts as an attention map, gating x_t^l to emphasize regions of longitudinal change. This lightweight operation runs at all

resolutions without architectural overhead. The temporally-informed features $\{x'^{l}_{t}\}$ are passed to the decoder to generate the follow-up segmentation. To our knowledge, this is the first framework to combine early prompt fusion with latent temporal fusion, applying each mechanism precisely where it is most effective.

Synthetic Longitudinal Pretraining. Real annotated multi-timepoint data is scarce, and without sufficient training data even a longitudinal architecture can collapse to cross-sectional behavior, ignoring the baseline scan entirely. We address this by extending the synthetic corpus of [15] to the promptable setting: 2,606 CT pairs with synthetic follow-ups generated via anatomy-informed deformation fields [11] simulating tumor growth, shrinkage, and acquisition variability are paired with full prompt simulation for both timepoints.

Prompt Simulation. To instill robustness against imprecise user interactions and registration errors, training prompts are generated via a 50/50 split: half are sampled from the ground-truth mask with probabilities weighted by $1/d^2$, where d is the distance to the centroid, and half are derived from the registered follow-up point, which may even fall outside the lesion.

3 Datasets

autoPET/CT IV Dataset. This dataset comprises longitudinal whole-body CT scans from 285 melanoma patients undergoing therapy response assessment [12] totalling 670 images. Each patient has at least one baseline and follow-up scan acquired during portal-venous phase on multiple Siemens scanners with standardized protocols. Two radiologists manually segmented all tumor lesions across timepoints with side-by-side verification to establish correspondence. The dataset includes challenging scenarios (splitting, merging, vanishing lesions) and provides pre-computed lesion centers at both timepoints, with registration-based propagated clicks simulating the *Verified Tracking* workflow.

The PanTrack Dataset. To evaluate generalization to a different anatomical domain, we curated *PanTrack*, comprising 45 patients with pancreatic adenocarcinoma amounting to 161 CT examinations (2–11 per patient; mean 3.6). All scans were acquired on identical Siemens protocols (portal-venous phase) at a single institution. An experienced radiologist with pancreatic imaging expertise manually segmented all pancreatic lesions across timepoints, including hepatic metastases if present. The cohort represents diverse trajectories: some patients underwent long-term stable chemotherapy, others showed rapid progression. Unlike melanoma metastases, pancreatic lesions exhibit fuzzy boundaries and subtle soft-tissue contrast, providing complementary evaluation under different radiological characteristics. This dataset is publicly released upon acceptance.

4 Experiments and Results

4.1 Experimental Setup

We develop and train our model exclusively on the autoPET IV dataset [12], optimizing a combined Dice and cross-entropy loss via SGD for 1000 epochs. We randomly reserve one-third of the patients as a held-out test set, splitting the remaining cohort into an 80/20 training and validation split. All architectural decisions and hyperparameters were fixed prior to evaluation on the held-out test set and PanTrack. Crucially, PanTrack was completely excluded from any form of training or model selection, serving as a rigorous, entirely unseen out-of-distribution (OOD) benchmark.

Evaluation Protocol. We evaluate under two tracking paradigms reflecting distinct clinical workflows. In the **Automatic Tracking** setting, only the baseline prompt p_0 is provided; the follow-up prompt $\hat{p}_t$ is generated fully automatically via uniGradICON registration. In the **Verified Tracking** setting, the ground-truth follow-up lesion centroid is provided, simulating a workflow where a clinician has accepted or swiftly corrected the registration-proposed location. All models receive prompts appropriate to the respective paradigm. Importantly, verified tracking eliminates catastrophic retrieval failures by design; remaining errors are purely delineation errors, which we quantify via Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD). We additionally report the lesion detection rate (LDR) to assess detection validity.

Baselines. For automatic tracking, we compare against a registration-based decoupled pipeline (Hering et al. [6], reimplemented with uniGradICON), an end-to-end tracker (LesionLocator [15]), and nnInteractive [8] prompted with the registration-propagated center. For verified tracking, we compare against interactive foundation models: nnInteractive [8], SegVol [4], and the official Universal Lesion Segmentation (ULS) model [2]. While off-the-shelf foundation models are evaluated zero-shot on autoPET, potentially giving our trained model an in-domain advantage, the PanTrack dataset provides a strictly fair, zero-shot evaluation ground for all methods.

4.2 Model Development: Unlocking the Longitudinal Prior

We perform a systematic ablation study on the autoPET IV validation set to isolate the contributions of our architectural design and training strategy (Tab. 1).

The Failure of Naive Longitudinal Fusion. Theoretically, providing the baseline scan should give the network strictly more information to resolve ambiguous boundaries. However, when trained from scratch, the naive longitudinal early fusion model (54.0 DSC) actually underperforms the single-timepoint baseline (55.8 DSC), confirming simply concatenating the inputs is insufficient.

Pretraining Activates Longitudinal Functionality. Introducing large-scale synthetic longitudinal pretraining [15] fundamentally changes the network's behavior. Pretraining largely prevents the cross-sectional collapse, allowing the

Table 1. Ablation study. Naive longitudinal concatenation from scratch (row 4) actually underperforms the single-timepoint baseline (row 1). While synthetic pretraining activates the architecture, Difference Weighting (DW) is essential to fully exploit the temporal prior. Best results in **bold**, second-best underlined.

Architecture	Temporal Fusion	Prompt Sim.	Pre-train	DSC ↑	NSD ↑	LDR ↑
Single Timepoint	–	✓	–	55.8	69.6	76.1
	–	✓	✓	<u>56.9</u>	69.1	75.8
Longitudinal	Concat.	–	–	45.8	59.0	66.2
	Concat.	✓	–	54.0	67.0	76.0
	Concat.	✓	✓	56.5	<u>70.6</u>	<u>77.7</u>
	Diff. Weighting	✓	✓	**58.5**	**72.3**	**78.8**

naive longitudinal model to finally surpass the single-timepoint baseline in boundary accuracy (70.6 vs. 69.1 NSD) and detection rate (77.7 vs. 75.8 LDR), proving that synthetic priors are essential for learning temporal correspondences.

Difference Weighting Maximizes the Temporal Prior. While pretraining activates the architecture, naive channel-wise concatenation remains a suboptimal fusion strategy. Replacing it with Difference Weighting (DW) explicitly forces the model to attend to longitudinal changes by computing normalized feature differences in latent space. This combined approach, i.e. synthetic pretraining to prevent collapse, and DW to provide the correct structural inductive bias, yields a decisive performance leap (58.5 DSC, 72.3 NSD), fully unlocking the value of longitudinal context.

Robustness via Prompt Simulation. Finally, we note that training strictly on perfect center-point prompts causes catastrophic degradation on realistic, slightly off-center registration prompts (45.8 DSC). Dynamically sampling simulated prompts during training is a critical requirement to ensure the localization robustness demanded by the *Verified Tracking* paradigm.

4.3 Comparison with State-of-the-Art

We evaluate our final model on both the autoPET IV test set and the OOD PanTrack dataset. Table 2 details the results against established baselines.

Automatic Tracking. While designed for human-in-the-loop verification, our method establishes state-of-the-art performance even fully automatically. The sharp decline of nnInteractive (43.3 DSC) highlights its vulnerability to off-center prompts. Conversely, our model is highly robust to residual registration errors, outperforming both decoupled pipelines (Hering et al.) and end-to-end trackers (LesionLocator) without requiring human intervention.

Verified Tracking. Clinician verification improves all methods, confirming that localization is the primary tracking bottleneck. Given identical verified prompts, our model achieves peak performance on both test sets (73.7 autoPET / 60.0

Table 2. Comparison on held-out test sets. Methods are grouped by tracking paradigm: automatic (prompted via registration) and verified (prompted via centroid). **Bold** indicates the best result per dataset with mean ±std obtained via bootstrapping.

		autoPET IV (test)			PanTrack (ours)		
	Method	DSC ↑	NSD ↑	LDR ↑	DSC ↑	NSD ↑	LDR ↑
Automatic	Hering et al. [6]	56.6±2.4	65.9±2.7	76.7±2.9	49.9±3.2	41.3±2.9	83.1±4.7
	nnInteractive [8]	43.3±2.7	49.5±3.0	61.3±3.4	34.3±3.5	27.2±2.7	59.3±5.7
	LesionLocator [15]	44.1±2.9	51.7±3.2	61.1±3.7	51.7±3.4	41.3±3.3	84.0±4.2
	Ours	**60.7**±2.6	**69.5**±2.9	**77.7**±3.2	**58.2**±3.0	**48.6**±3.0	**93.7**±3.0
Verified	SegVol [4]	45.8±2.0	57.5±2.0	79.4±2.8	43.7±2.3	33.4±2.6	87.4±4.2
	ULS Model [2]	68.7±1.6	79.7±1.9	93.3±1.7	53.4±2.7	42.8±2.9	**95.5**±2.4
	nnInteractive [8]	59.6±2.4	66.6±2.7	80.6±2.8	52.8±3.1	42.2±2.9	93.5±2.9
	Ours	**73.7**±1.5	**84.0**±1.7	**95.5**±1.2	**60.0**±2.8	**49.7**±2.9	94.7±2.7

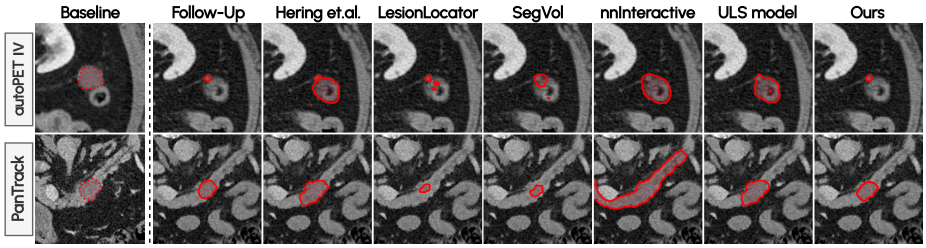

Fig. 2. Qualitative comparison on autoPET IV (top) and PanTrack (bottom). Single-timepoint baselines struggle with ambiguous lesion volumes. In the top row, a shrinking lesion borders the colon; without the baseline appearance as context, competing models fail to isolate the correct structure when prompted near the organ boundary.

PanTrack DSC), substantially outperforming prior promptable models. This margin isolates the value of the longitudinal prior: our model can leverage the baseline appearance to delineate ambiguous boundaries that single-timepoint models fail to resolve (see Fig. 2).

OOD Generalization on PanTrack. Despite PanTrack's different scanners, institutions, and fuzzy lesion boundaries, our method maintains its superiority. Strikingly, our model's *automatic* tracking performance (58.2 DSC) not only closely approaches our *verified* performance (60.0 DSC), but outright surpasses the *verified* results of all competing models (e.g., ULS at 53.4 DSC). This proves our framework delivers highly reliable automated tracking "in the wild" while natively preserving the safety of optional clinician correction.

5 Conclusion and Outlook

We present a robust framework for longitudinally-informed lesion tracking that secured first place in the MICCAI autoPET IV challenge and demonstrates

state-of-the-art generalization on *PanTrack*, a novel out-of-distribution dataset. We propose "Verified Tracking" as a clinically viable middle ground for standard RECIST 1.1 workflows. By requiring clinicians to merely verify or correct a single follow-up point, this paradigm averts catastrophic retrieval failures and gracefully handles complex topological changes like splitting or vanishing lesions. Once anchored, our architecture leverages explicit temporal fusion and synthetic pretraining to fully exploit the baseline appearance, significantly improving delineation accuracy. While our framework currently assumes a known baseline lesion, a step readily automated by off-the-shelf detectors, it successfully isolates and solves the critical bottleneck of longitudinal correspondence. Building on our strong zero-shot OOD performance, future work will focus on prospective clinical reader studies and exploring richer prompt modalities beyond points, such as free-text descriptions of lesion characteristics [16]. To foster further research in generalizable tracking, we publicly release the *PanTrack* dataset, along with our code and model weights.

Acknowledgments. This work was partly funded by the Helmholtz Information and Datascience School (HIDSS) and Helmholtz Imaging (HI), platforms of the Helmholtz Incubator on Information and Data Science. Supported by the Helmholtz Foundation Model Initiative (HFMI) through the pilot project THRP (The Human Radiome Project). Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 402688427. This project was funded within the DKTK Heidelberg Seed Funding 25 program. M.R. is funded through a Google PhD Fellowship.

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

References

1. Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R.L., Soerjomataram, I., Jemal, A.: Global cancer statistics 2022: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians **74**(3), 229–263 (2024)
2. de Grauw, M., Scholten, E., Smit, E., Rutten, M., Prokop, M., van Ginneken, B., Hering, A.: The uls23 challenge: A baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. Medical Image Analysis **102**, 103525 (2025). https://doi.org/https://doi.org/10.1016/j.media.2025.103525, https://www.sciencedirect.com/science/article/pii/S1361841525000738
3. Denner, S., Khakzar, A., Sajid, M., Saleh, M., Spiclin, Z., Kim, S.T., Navab, N.: Spatio-temporal learning from longitudinal data for multiple sclerosis lesion segmentation. In: International MICCAI Brainlesion Workshop. pp. 111–121. Springer (2020)
4. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. Advances in Neural Information Processing Systems **37**, 110746–110783 (2024)
5. Eisenhauer, E.A., Therasse, P., Bogaerts, J., Schwartz, L.H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., et al.: New response evaluation criteria in solid tumours: revised recist guideline (version 1.1). European journal of cancer **45**(2), 228–247 (2009)

6. Hering, A., Peisen, F., Amaral, T., Gatidis, S., Eigentler, T., Othman, A., Moltz, J.H.: Whole-body soft-tissue lesion tracking and segmentation in longitudinal ct imaging studies. In: Medical Imaging with Deep Learning. pp. 312–326. PMLR (2021)
7. Hering, A., Westphal, M., Gerken, A., Almansour, H., Maurer, M., Geisler, B., Kohlbrandt, T., Eigentler, T., Amaral, T., Lessmann, N., et al.: Improving assessment of lesions in longitudinal ct scans: a bi-institutional reader study on an ai-assisted registration and volumetric segmentation workflow. International Journal of Computer Assisted Radiology and Surgery **19**(9), 1689–1697 (2024)
8. 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 (2025), https://arxiv.org/abs/2503.08373
9. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jäger, P.F.: nnU-Net Revisited: A Call for Rigorous Validation in 3D Medical Image Segmentation . In: proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. vol. LNCS 15009. Springer Nature Switzerland (October 2024)
10. Jacobs, C., Schreuder, A., van Riel, S.J., Scholten, E.T., Wittenberg, R., Wille, M.M.W., de Hoop, B., Sprengers, R., Mets, O.M., Geurts, B., et al.: Assisted versus manual interpretation of low-dose ct scans for lung cancer screening: impact on lung-rads agreement. Radiology: Imaging Cancer **3**(5), e200160 (2021)
11. Kovacs, B., Netzer, N., Baumgartner, M., Eith, C., Bounias, D., Meinzer, C., Jäger, P.F., Zhang, K.S., Floca, R., Schrader, A., et al.: Anatomy-informed data augmentation for enhanced prostate cancer detection. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 531–540. Springer (2023)
12. Küstner, T., Peisen, F., Gatidis, S., Wagner, A., Megne, O., Othman, A., Sanner, A., Loß au, T., Moltz, J.H., Kohlbrandt, T., Hering, A.: Longitudinal-ct. https://fdat.uni-tuebingen.de/records/qwsry-7t837 (Mar 2025). https://doi.org/10.57754/FDAT.qwsry-7t837, version v1, Published March 16, 2025
13. Lu, S.L., Xiao, F.R., Cheng, J.C.H., Yang, W.C., Cheng, Y.H., Chang, Y.C., Lin, J.Y., Liang, C.H., Lu, J.T., Chen, Y.F., et al.: Randomized multi-reader evaluation of automated detection and segmentation of brain tumors in stereotactic radiosurgery with deep neural networks. Neuro-oncology **23**(9), 1560–1568 (2021)
14. Rocholl, N., Smit, E., Prokop, M., Hering, A.: Unstable prompts, unreliable segmentations: A challenge for longitudinal lesion analysis. arXiv preprint arXiv:2507.19230 (2025)
15. Rokuss, M., Kirchhoff, Y., Akbal, S., Kovacs, B., Roy, S., Ulrich, C., Wald, T., Rotkopf, L.T., Schlemmer, H.P., Maier-Hein, K.: Lesionlocator: Zero-shot universal tumor segmentation and tracking in 3d whole-body imaging. In: Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR). pp. 30872–30885 (June 2025)
16. Rokuss, M., Langenberg, M., Kirchhoff, Y., Isensee, F., Hamm, B., Ulrich, C., Regnery, S., Bauer, L., Katsigiannopulos, E., Norajitra, T., Maier-Hein, K.: Voxtell: Free-text promptable universal 3d medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2026)
17. Rokuss, M.R., Kirchhoff, Y., Roy, S., Kovacs, B., Ulrich, C., Wald, T., Zenk, M., Denner, S., Isensee, F., Vollmuth, P., Kleesiek, J., Maier-Hein, K.: Longitudinal

segmentation of ms lesions via temporal difference weighting. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 64–74. Springer (2024)
18. Szeskin, A., Rochman, S., Weiss, S., Lederman, R., Sosna, J., Joskowicz, L.: Liver lesion changes analysis in longitudinal cect scans by simultaneous deep learning voxel classification with simu-net. Medical Image Analysis **83**, 102675 (2023)
19. Tian, L., Greer, H., Kwitt, R., Vialard, F.X., San José Estépar, R., Bouix, S., Rushmore, R., Niethammer, M.: unigradicon: A foundation model for medical image registration. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 749–760. Springer (2024)
20. Vizitiu, A., Mohaiu, A.T., Popdan, I.M., Balachandran, A., Ghesu, F.C., Comaniciu, D.: Multi-scale self-supervised learning for longitudinal lesion tracking with optional supervision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 573–582. Springer (2023)
21. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: Scribbleprompt: Fast and flexible interactive segmentation for any biomedical image (2024), `https://arxiv.org/abs/2312.07381`
22. Wu, Y., Wu, Z., Shi, H., Picker, B., Chong, W., Cai, J.: Coactseg: Learning from heterogeneous data for new multiple sclerosis lesion segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 3–13. Springer (2023)
23. 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)