Promptable Longitudinal Lesion Segmentation in Whole-Body CT

Yannick Kirchhoff[1,2,3], Maximilian Rokuss[1,2,3], Fabian Isensee[1,4], and Klaus H. Maier-Hein[1,4,5]

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

Abstract. Accurate segmentation of lesions in longitudinal whole-body CT is essential for monitoring disease progression and treatment response. While automated methods benefit from incorporating longitudinal information, they remain limited in their ability to consistently track individual lesions across time. Task 2 of the autoPET/CT IV Challenge addresses this by providing lesion localizations and baseline delineations, framing the problem as longitudinal promptable segmentation. In this work, we extend the recently proposed LongiSeg framework with promptable capabilities, enabling lesion-specific tracking through point and mask interactions. To address the limited size of the provided training set, we leverage large-scale pretraining on a synthetic longitudinal CT dataset. Our experiments show that pretraining substantially improves the ability to exploit longitudinal context, yielding an improvement of up to 6 Dice points compared to models trained from scratch. These findings demonstrate the effectiveness of combining longitudinal context with interactive prompting for robust lesion tracking. Code is publicly available at https://github.com/MIC-DKFZ/LongiSeg/tree/autoPET.

Keywords: autoPET/CT IV Challenge · Longitudinal Imaging · Interactive Segmentation · Lesion Tracking.

1 Introduction

Accurate lesion segmentation in whole-body CT is crucial for reliable monitoring of tumor progression and treatment response in oncology. While deep learning-based methods show strong performance in cross-sectional lesion segmentation [1], tracking individual lesions across timepoints remains a challenging task due to anatomical changes, differences in patient positioning, heterogeneity

of lesion appearances, and typically small publicly available datasets.

The autoPET/CT IV challenge directly addresses these challenges. In Task 2, the organizers provide a longitudinal whole-body CT dataset [6], consisting of 300 patients with baseline and follow-up images, accompanied by ground truth segmentation masks and localization information for each individual lesion. The challenge effectively poses the task of lesion tracking as a longitudinal promptable segmentation problem.

This formulation aligns with recent advances in medical image segmentation. First, after the introduction of the Segment Anything Model (SAM) [4], interactive methods also gained traction in the medical domain [10,9,2]. Second, longitudinal frameworks such as *LongiSeg* [8] and *LesionLocator* [7] have demonstrated the benefits of incorporating temporal context for longitudinal image segmentation and lesion tracking.

In this work, we build upon these developments and propose a framework tailored to Task 2 of the autoPET/CT IV Challenge. Specifically, we extend the LongiSeg framework with promptable capabilities for point- and mask-based interactions. Furthermore, we employ large-scale pretraining on a synthetic longitudinal CT dataset to overcome limitations of the challenge dataset size. We systematically evaluate different design decisions, including model input, prompt representation and pretraining strategies, demonstrating the benefits of longitudinal pretraining for robust lesion tracking.

2 Datasets

2.1 Training data

The model is trained on the provided dataset [6]. This dataset includes baseline and follow-up scans of 300 patients, including ground truth segmentations, lesion centers and propagated center locations in the follow-up for each individual lesion. Out of the 300 provided patients, 15 were excluded due to points being at the edge of scans, which likely represents cases, where the lesion location is not present on the follow-up.

2.2 Pretraining data

In addition to the provided dataset we utilized the *LesionLocator* syntetic longitudinal CT dataset [7] for pretraining of our model. This dataset includes real CT volumes of 2625 patients which are augmented using anatomy informed data augmentation [5] to generate a synthetic baseline scan.

3 Methods

Our approach builds upon the recently proposed *LongiSeg* framework [8], which was originally developed for fully automatic longitudinal segmentation of aligned baseline–follow-up scans. In contrast, the more recent *LesionLocator* [7] is a promptable framework that enables lesion tracking starting from a single baseline prompt by propagating it to future timepoints. While LesionLocator is well suited for the general setting of lesion tracking without follow-up annotations, the autoPET/CT IV challenge provides prompts in both baseline and follow-up scans. This renders prompt propagation unnecessary and makes the LongiSeg formulation, with direct use of predefined patches around the provided prompts, the more appropriate choice for this task.

3.1 Network architecture

Backbone and longitudinal inputs: We utilize the powerful ResEncL [3] preset as a model backbone, which generates a deep U-Net architecture with multiple residual blocks at each encoder layer. Longitudinal inputs are aligned via the given center locations and concatenated along the channel dimension as input to the model.

Prompt representation: We follow the same strategy we use for longitudinal inputs of providing both point as well as mask prompts via additional input channels to the U-Net. Point prompts are represented as Gaussian blobs, which we rescale to unit intensity at the center to approximately match the intensity normalization of other input channels. Initial experiments showed significant gains with respect to normalizing the Gaussians to unit volumes.

3.2 Data sampling:

During training single lesions are sampled from random patients. Baseline and follow-up scans are aligned via the given center locations and both scans are randomly shifted by up to 4 voxels in each direction. Patches from both scans are extracted, placing the center at a random location in the inner half of each patch. During inference the center locations are perfectly aligned and patches are extracted such that the center lies directly in the middle of the patches to guarantee best possible performance.

3.3 Inference

During inference the network does a single forward pass per lesion and either exports each prediction separately or merges the predictions into a single multi-label segmentation map.

Table 1: Results from the five-fold cross-validation. Predictions are merged according to the ground truths and metrics are calculated for each merged lesion pair and averaged over patients. Fold 4 is excluded from the mean as it was unstable during training, full results can be found in the appendix 2.

Setting	Dice↑	FNvol↓	FPvol↓
Cross Sectional + Point	53.06	1532	**113**
Cross Sectional + Mask	56.64	808	195
Longitudinal + Mask + Point	55.78	1005	355
Longitudinal Batch Size 2	58.08	736	374
Pretrained nnInteractive Weights	59.28	763	630
Pretrained LesionLocator Weights	59.57	689	626
Pretrained Cross Sectional + Mask	62.27	**266**	372
Pretrained Synth. Longitudinal Data	62.89	366	177
Final	**63.71**	343	144

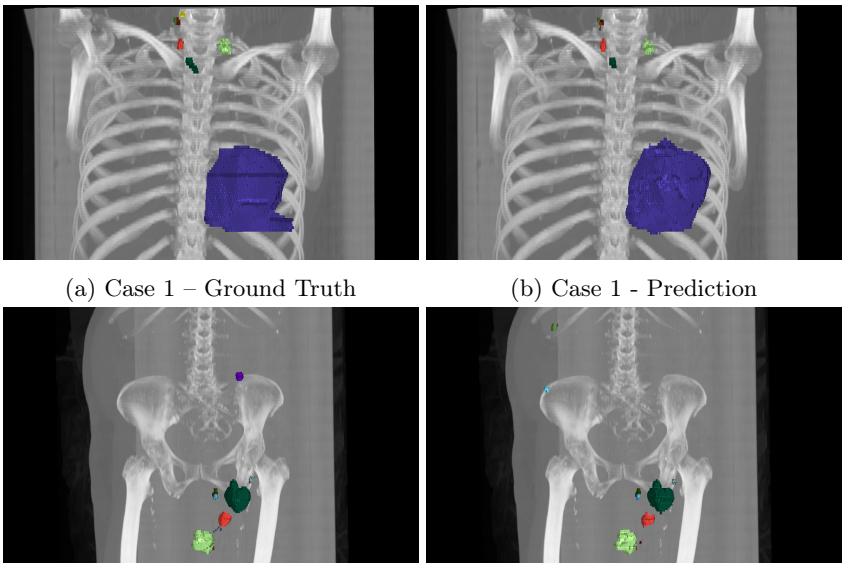
4 Results

Table 1 reports the results of our five-fold cross-validation experiments, evaluating Dice, false negative volume (FNvol) and false positive volume (FPvol). Different ablations are shown, including different inputs, batch sizes and pretrainings. Without pretraining, the model trained only on the current image and the prior mask performs best, indicating that it is not able to learn the full longitudinal context from the provided dataset alone. However, pretraining the models on the large synthetic dataset enables the model to properly utilize the information from the previous timepoint, outperforming the single timepoint solution by 0.6 dice points. Experiments with different batch sizes show that a smaller batch size of 2 performs noticeably better than our default batch size of 4. While an initialization with the pretrained LesionLocator [7] checkpoint performs slightly better than using nnInteractive [2] as an initialization, both do not come close to longitudinal pretraining on the synthetic dataset. Scaling up this pretraining with a larger batch size gives a further boost in performance, improving the dice score, while simultaneously reducing both false negative and false positive volumes. Figure 1 shows qualitative results of our best model on two cases from the cross-validation, underlining its strong performance.

Test Set Submission: For the final submission we ensemble the five folds from the best performing model on the cross-validation. This is pretrained with a large batch size on the synthetic longitudinal dataset and gets both timepoints, the prior segmentation and the point prompt as inputs via channel concatenation.

5 Conclusion

This paper presents our contribution to Task 2 of the autoPET/CT IV. We extended the LongiSeg framework to incorporate point- and mask-based prompts



(a) Case 1 – Ground Truth (b) Case 1 - Prediction

(c) Case 2 - Ground Truth (d) Case 2 - Prediction

Fig. 1: Qualitative results of our final model on two cases from the cross-validation. The top row shows Case 1 (ID: `013d407166`), and the bottom row shows Case 2 (ID: `31fefc0e57`). Each pair displays the ground truth (left) and model prediction (right). The model tracks and segments most lesions accurately, with inaccuracies mainly in boundary regions.

to enable tracking lesions across different timepoints. Additionally, we use a large synthetic longitudinal dataset to pretrain our model, which significantly improves performance upon training from scratch. Our final submission is based on a five-fold ensemble with large-scale pretraining on the synthetic dataset. Our results demonstrate the effectiveness of incorporating longitudinal information together with point- and mask-based prompts within the LongiSeg framework for efficient longitudinal lesion tracking.

Acknowledgements The present contribution is supported by the Helmholtz Association under the joint research school "HIDSS4Health – Helmholtz Information and Data Science School for Health". This work was partly funded by Helmholtz Imaging (HI), a platform of the Helmholtz Incubator on Information and Data Science.

References

1. Isensee, F., Jaeger, P.F., Kohl, S.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 (2021)
2. Isensee, F., Rokuss, M., Krämer, L., Dinkelacker, S., Ravindran, A., Stritzke, F., Hamm, B., Wald, T., Langenberg, M., Ulrich, C., et al.: nninteractive: Redefining 3d promptable segmentation. arXiv preprint arXiv:2503.08373 (2025)
3. Isensee, F., Wald, T., Ulrich, C., Baumgartner, M., Roy, S., Maier-Hein, K., Jaeger, P.F.: nnu-net revisited: A call for rigorous validation in 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 488–498. Springer (2024)
4. 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: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
5. 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)
6. 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
7. 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. pp. 30872–30885 (2025)
8. Rokuss, M.R., Kirchhoff, Y., Roy, S., Kovacs, B., Ulrich, C., Wald, T., Zenk, M., Denner, S., Isensee, F., Vollmuth, P., et al.: 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)

9. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., et al.: Sam-med3d: towards general-purpose segmentation models for volumetric medical images. In: European Conference on Computer Vision. pp. 51–67. Springer (2024)
10. Wong, H.E., Rakic, M., Guttag, J., Dalca, A.V.: Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In: European Conference on Computer Vision. pp. 207–229. Springer (2024)

A Full model results

Table 2: Results from the five-fold cross-validation. Predictions are merged according to the ground truths and metrics are calculated for each merged lesion pair and averaged over patients.
: Training on Fold 4 collapsed with resulting dice of 0

Setting	Dice↑	FNvol↓	FPvol↓
Cross Sectional + Point	52.81	1353	**107**
Cross Sectional + Mask	56.28	771	266
Longitudinal + Mask + Point	55.62	902	394
Longitudinal Batch Size 2*	46.47	2456	299
Pretrained nnInteractive	58.90	735	616
Pretrained LesionLocator	58.96	678	608
Pretrained Cross Sectional + Mask	61.20	**345**	346
Pretrained Longitudinal	61.94	404	243
Final	**62.49**	386	132

Table 3: Algorithm details

Team name	**algorithm name** (as submitted on grand-challenge)	**data pre-processing**	**data post-processing**
LesionLocator	LongiLesionLocator	normalization & resampling	-

training data augmentation	**test time augmentation**	**ensembling** (e.g. cross-validation, model ensemble, ...)	**standardized framework?** (e.g. nnUNet, MONAI, ...)
nnUNet augmentations, longitudinal misalignments, center jiggle	-	5-fold ensembling	LongiSeg

network architecture (e.g. UNet (3D))	**loss**	**training data**	**data/model dimensionality and size**
UNet (3D)	DSC + CE	autoPET IV, synthetic longitudinal dataset	3D: 112x224x224

use of pre-trained models	**GPU hardware for training**
yes, pretraining on synthetic dataset	Nvidia A100