

nnInteractive: Redefining 3D Promptable Segmentation

Fabian Isensee^{1,4*}, Maximilian Rokuss^{1,2*}, Lars Krämer^{1,4*}, Stefan Dinkelacker¹, Ashis Ravindran¹, Florian Stritzke⁵, Benjamin Hamm^{1,3}, Tassilo Wald^{1,2,4}, Moritz Langenberg^{1,2}, Constantin Ulrich¹, Jonathan Deissler^{1,2}, Ralf Floca¹, Klaus Maier-Hein^{1,2,3,4,6,7}

¹German Cancer Research Center, Division of Medical Image Computing, Germany

²Faculty of Mathematics and Computer Science and ³Medical Faculty - Heidelberg University
⁴Helmholtz Imaging, ⁵Department of Radiation Oncology, Heidelberg University Hospital, Germany
⁶HIDSS4Health, Heidelberg ⁷Pattern Analysis and Learning Group, Heidelberg University Hospital

{f.isensee, maximilian.rokuss}@dkfz-heidelberg.de

Abstract

Accurate and efficient 3D segmentation is essential for both clinical and research applications. While foundation models like SAM have revolutionized interactive segmentation, their 2D design and domain shift limitations make them ill-suited for 3D medical images. Current adaptations address some of these challenges but remain limited, either lacking volumetric awareness, offering restricted interactivity, or supporting only a small set of structures and modalities. Usability also remains a challenge, as current tools are rarely integrated into established imaging platforms and often rely on cumbersome web-based interfaces with restricted functionality. We introduce nnInteractive, the first comprehensive 3D interactive open-set segmentation method. It supports diverse prompts—including points, scribbles, boxes, and a novel lasso prompt—while leveraging intuitive 2D interactions to generate full 3D segmentations. Trained on 120+ diverse volumetric 3D datasets (CT, MRI, PET, 3D Microscopy, etc.), nnInteractive sets a new state-of-the-art in accuracy, adaptability, and usability. Crucially, it is the first method integrated into widely used image viewers (e.g., Napari, MITK), ensuring broad accessibility for real-world clinical and research applications. Extensive benchmarking demonstrates that nnInteractive far surpasses existing methods, setting a new standard for AI-driven interactive 3D segmentation. nnInteractive is publicly available: [Napari plugin](#), [MITK integration](#), [Python backend](#).

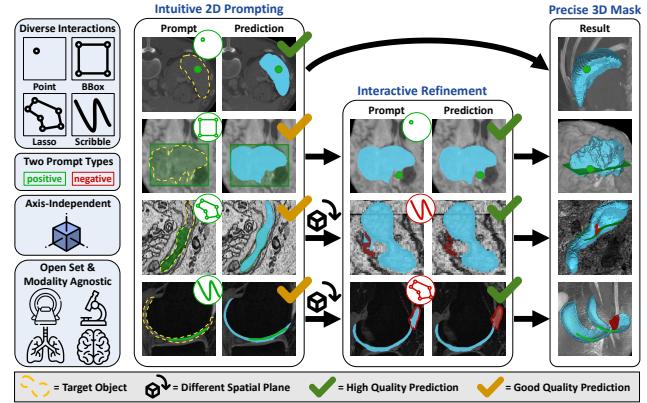


Figure 1. **nnInteractive** fully unlocks the potential of 3D interactive segmentation. Supporting a diverse set of prompting styles, it generates full 3D segmentations from intuitive 2D interactions. Prompts can be arbitrarily mixed and placed on any axis. nnInteractive is open set and supports all modalities. It quickly adapts to user input to accurately segment any target structure.

1. Introduction

Precise and efficient 3D segmentation is essential across fields such as medical imaging, biology, and industrial inspection, where volumetric data provides critical insights. Unlike 2D segmentation, which processes individual images, 3D segmentation must maintain volumetric consistency while handling high-dimensional data efficiently. This requires specialized methods that can adapt to diverse structures, imaging modalities, and real-world constraints. While fully automated segmentation models have achieved remarkable performance in specific tasks [44, 46, 47, 122, 134], their reliance on predefined training labels and distributions limits generalization to unseen structures and domains. As a result, interactive segmentation has gained traction

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

tion as a way to integrate user guidance, enhancing flexibility, accuracy, and real-world applicability.

Recent advances in foundation models, particularly interactive vision models like SAM [60, 112], enable general-purpose segmentation across diverse datasets via prompt-based interactions. By using simple inputs like points and bounding boxes these models reduce the need for task-specific training. SAM’s direct application to medical imaging has shown promise [121], which has driven a plethora of advancements in adapting interactive segmentation methods to the medical domain. Existing models [24, 88, 144], however, primarily operate on 2D slices, failing to account for the volumetric nature of medical scans. This leads to inconsistencies across slices and necessitates labor-intensive manual refinements when extending segmentations to full 3D volumes. Furthermore, models that do support volumetric segmentation are often limited to single imaging modalities such as computed tomography (CT) [30, 42, 116] or are constrained to closed-set segmentation, meaning they can only identify structures seen during training [42, 70]. These limitations hinder their generalizability and practical usability, particularly in clinical workflows that demand adaptability to new and unseen structures and image acquisition protocols. Beyond these challenges, current interactive segmentation frameworks also suffer from restricted interaction paradigms. Most rely on simple prompts such as points and boxes [38, 88], with few methods supporting more intuitive inputs like scribbles [144]. The lack of user-friendly interactions further complicates usability, as existing tools often require cumbersome 3D bounding box annotations rather than leveraging natural 2D interactions for volumetric segmentation. Additionally, some methods do not support negative prompts [30] or interactive refinement [88], both of which are essential for practical segmentation workflows. Finally, usability concerns persist, as existing solutions lack user-friendly interfaces integrated into established annotation platforms, limiting their adoption in real-world applications.

In this work, we introduce nnInteractive, a 3D interactive segmentation framework that systematically addresses key challenges in volumetric annotation. Rather than proposing new architectural innovations, our approach prioritizes usability, interaction diversity, positive and negative prompting, and computational efficiency. To ensure broad applicability and generalization, our model is trained on an unprecedentedly large and multimodal 3D dataset comprising over 120 publicly available datasets with 64,518 volumes spanning multiple imaging modalities and anatomical structures. Beyond its technical capabilities, it is designed for seamless real-world adoption. It is integrated into widely used annotation platforms such as Napari [127] and MITK Workbench [100], providing an accessible and efficient tool for both research and clinical workflows. Ex-

tensive evaluation across diverse and out-of-distribution datasets demonstrates that nnInteractive establishes a new benchmark in 3D interactive segmentation, combining state-of-the-art performance with practical usability. We present the following *key contributions*:

- nnInteractive presents the first 3D interactive open-set segmentation model, supporting a wide range of positive and negative prompts (points, scribbles, bounding boxes, and lasso) across multiple imaging modalities (CT, MR, PET, etc.). It enables full 3D segmentation from intuitive 2D interactions, with an adaptive AutoZoom mechanism for large structures. It builds on nnU-Net’s best practices of carefully curated optimal design choices rather than a new novel architecture.
- Extensive benchmarking on 14 datasets shows superior segmentation accuracy, interactive refinement, and clinical usability, with the proposed lasso interaction providing the best guidance signal to the model.
- It is trained on an unprecedented scale with over 120 diverse 3D datasets, including a wide range of modalities, anatomical structures and label variations as well as novel SuperVoxels. Natural and simulated label variations enable nnInteractive to resolve segmentation ambiguities based on user intent.
- Optimized for real-world usability with <10 GB VRAM requirement, rapid inference and seamless integration into Napari and MITK Workbench.

2. Method

We design nnInteractive around three key principles: usability, interaction diversity, generalization and computational efficiency. This section details its network architecture, the transformation of 2D prompts into 3D masks, supported interaction types, user simulation, strategies to handle segmentation ambiguities, and our proposed Auto Zoom to predict targets that exceed the patch size of the model. Further details can be found in Appendix A1.

2.1. Network Architecture

Despite the widespread adoption of Transformer-based models in 2D computer vision, UNet architectures continue to dominate 3D medical image segmentation, consistently delivering state-of-the-art performance in benchmarks and competitions [11, 27, 29, 35, 47, 148]. Given these advantages, we adopt a UNet-based design over Transformer alternatives and build upon the nnU-Net framework [46] employing the Residual Encoder (ResEnc-L) configuration [47] as our backbone. A key distinction of our approach lies in how prompts are incorporated. Existing interactive segmentation models [24, 30, 60, 88, 112, 140] typically encode images first and then integrate user-provided prompts in latent space. While this approach has been effective in 2D leveraging large pretrained models, the absence of proper

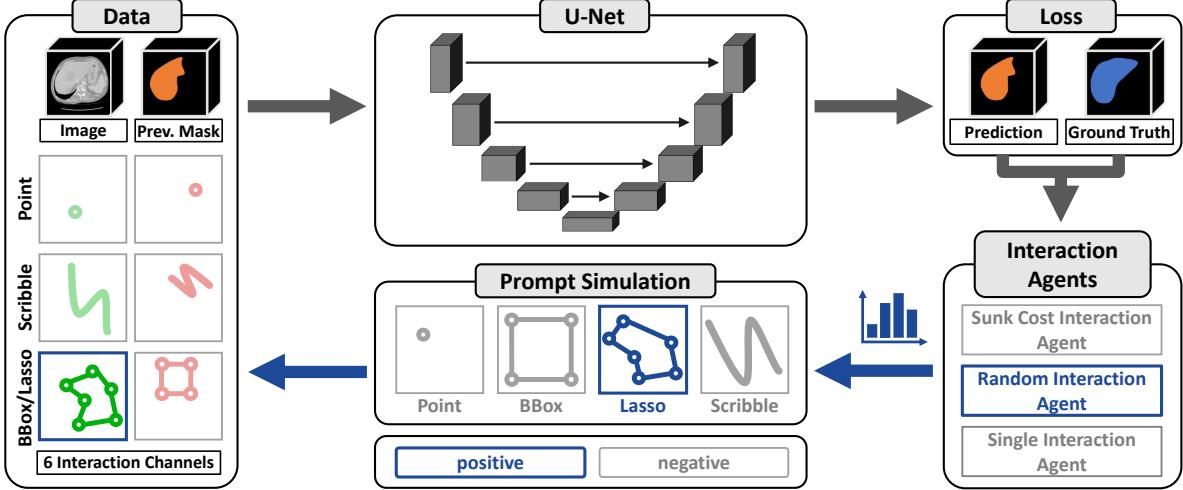


Figure 2. Overview of the nnInteractive Training Pipeline. The model first receives an input image and an initial prompt. The network then generates a prediction, which is used to compute the loss and identify false positive/negative areas. Based on the interaction agent simulating user input, a new prompt is sampled and added to the network input along with the current prediction.

3D foundation models renders this strategy suboptimal for 3D medical imaging. Instead, we adopt an early prompting strategy similar to [144] (see Fig. 2), where user inputs are directly incorporated as additional channels, ensuring prompts influence the entire feature extraction process. This allows the model to learn task-relevant representations from the highest resolution.

2.2. From 2D Prompt to 3D Mask

A core principle of nnInteractive is enhancing usability by bridging the gap between intuitive 2D annotation and full 3D segmentation. Modern 2D promptable models require separate prompts for each slice, and existing 3D approaches often rely on volumetric inputs such as 3D bounding boxes, which are challenging to define precisely from 2D views and can introduce segmentation errors due to excess empty space. nnInteractive efficiently predicts 3D masks from lower-dimensional prompts (see Fig. 1). Prompts can be either discrete points $p \in \mathbb{R}^3$ or structured 2D annotations $p \in \mathbb{R}^{m \times n}$ (e.g., scribbles, bounding boxes, lassos). These prompts can be placed on any plane, from which nnInteractive generates a complete 3D segmentation mask, significantly minimizing annotation effort.

2.3. Interaction Simulation

nnInteractive supports a comprehensive range of spatial prompt types, including points, bounding boxes, scribbles and lasso selections. Each prompt type is encoded in two separate input channels (positive and negative), see Fig. 2. Lasso and bounding boxes share a pair of input channels. Inspired by Photoshop’s selection tool, the lasso prompt provides a more precise alternative to bounding boxes while

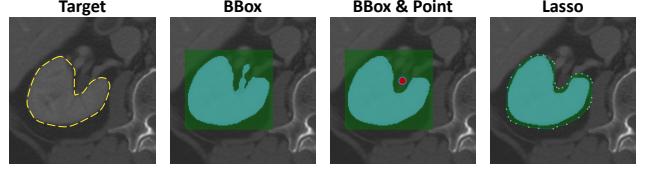


Figure 3. Bounding Box vs. Lasso. A bounding box interaction often requires additional refinement, whereas a fine-grained lasso interaction enables precise segmentation in a single step.

requiring comparable annotation effort. Unlike bounding boxes, which often enclose irrelevant structures and necessitate additional refinement (Fig. 3), lasso selections allow users to loosely outline the object without exact tracing. We now describe nnInteractive’s prompt and interaction generation process. During training, the initial prompt simulation is derived from the ground truth, while subsequent interactions are guided by the current prediction error. Since predicted masks often contain both false positives (FP) and false negatives (FN), a connected-component V is first selected from either a FP or FN region. For lasso, 2D bounding boxes, and scribbles, a representative slice S is sampled from V . The interaction is then simulated and assigned to the corresponding input channel. Upon adding a followup interaction, existing interactions are decayed by multiplying their intensity with 0.9. The network also receives its latest prediction as additional input.

Identifying Error Regions. Error regions are identified by computing false positive (FP) and false negative (FN) areas from the difference between the ground truth y and prediction \hat{y} . A random error component V is then se-

lected with probability proportional to its size. For bounding boxes and lassos, additional fragmentation is applied to prevent large, thin border regions—caused by slight over- or under-segmentation—from dominating the correction process. This is achieved by multiplying a thresholded Perlin noise mask, which breaks elongated structures before performing connected component analysis.

2D Slice Sampling. A 2D slice S is extracted from the selected 3D error component V . Slices are sampled probabilistically based on the foreground volume distribution across axial, sagittal, and coronal planes, with a bias toward slices containing more foreground voxels. The chosen slice S serves as the foundation for generating scribbles, lassos, and bounding box interactions.

Point Interactions. A Point interaction selects a representative location within the selected error component V . We compute the normalized Euclidean Distance Transform (EDT) D for all voxels $x \in V$, $D(x) \rightarrow [0, 1]$ with D assigning the highest values to central voxels. The point prompt location is then sampled using either a center-biased approach ($\alpha = 8$) or uniform sampling ($\alpha = 1$), with sampling probability:

$$p(x) = \frac{D(x)^\alpha}{\sum_{z \in V} D(z)^\alpha}.$$

The sampled point is expanded into a sphere which is then converted to a soft mask with maximum intensity at the center via an additional normalized EDT.

Bounding Box Interactions. Bounding boxes enclose the 2D error region in the selected slice S . To introduce variability, the bounding box is randomly augmented:

- **Jittering:** Each boundary is perturbed by an offset sampled from $\mathcal{U}(-0.05, 0.05) \cdot d$, where d is the size of the bounding box in that dimension.
- **Shifting:** The entire box is translated within the same range as jittering.
- **Scaling:** A factor $s \sim \mathcal{U}(0.8, 1.2)$ is applied per dimension, maintaining uniform scaling with probability 0.3.

Lasso Interactions. Lasso interactions provide a more precise alternative to bounding boxes, allowing users to loosely outline the object of interest. The lasso mask is generated in two steps:

1. **Coarse mask generation:** An enclosure of the 2D error mask is formed using closing and dilation, where structuring element sizes are adapted to the object shape via the directional Euclidean distance transform (EDT).
2. **Deformation:** A random displacement field is applied, with deformation magnitudes sampled proportionally to the directional EDT, ensuring realistic variability.

Scribble Interactions. Inspired by ScribblePrompt [144], nnInteractive generates three types of scribbles—center, line, and contour—each selected with equal probability. Unlike ScribblePrompt, which uses Perlin noise to break scribbles into multiple disconnected parts and thus regularly simulates multiple disconnected scribbles at once, our approach avoids such fragmentation by instead truncating scribbles at randomly sampled upper and lower coordinate bounds along each axis. Our implementation furthermore improves upon consistency by guaranteeing a fixed scribble width and improves parametrization by tying deformation parameters to object size.

- **Center Scribbles:** Extracting the skeleton of the 2D error mask, then truncating to simulate partial annotation.
- **Line Scribbles:** Connecting two random points from the 2D error mask.
- **Contour Scribbles:** First eroding the 2D error mask and then computing the truncated contour of the eroded object.

Each scribble undergoes structured deformation using a random displacement field, where deformation magnitudes are sampled proportionally to the directional Euclidean distance transform (EDT). Skeletonization and dilation are then applied to maintain consistent thickness.

2.4. User Simulation

While nnInteractive supports diverse prompting styles, real users tend to follow consistent patterns rather than switching randomly. Disregarding this could lead to overfitting on unrealistic prompt strategies that do not reflect real-world usage. To maximize generalization, we introduce *simulated user agents* that guide interaction sequences over multiple steps during training. Overall, we introduce three agents: i) The *Random agent*, selects a different prompting/interaction type at each iteration, representing a user who switches prompts at will. ii) The *Sunk Cost agent* represents a user that prefers one interaction type and sticks with it for several iterations before switching to another. We simulate this with a high probability of keeping the current prompting style and a low probability of randomly switching to a new one. iii) The *Single Interaction agent* represents a user that strongly prefers one type of interaction and keeps it throughout the entire refinement process.

2.5. Auto zoom

One key limitation of 3D models is the need for patch-wise processing due to VRAM constraints. Without a dedicated mechanism, objects larger than the patch size get truncated at patch borders, leading to incomplete segmentation. Existing 3D methods either ignore this issue or rely on additional interactions to prompt new patches [42, 140]. SegVol [30] addresses this by first segmenting at a coarse resolution, heavily downscaling images to fit a $32 \times 256 \times 256$ patch,

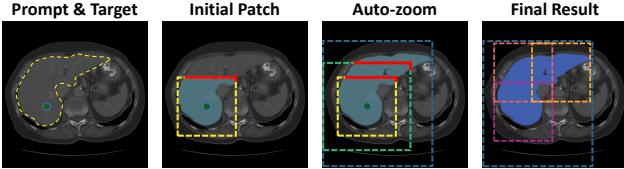


Figure 4. **Auto Zoom.** nnInteractive adaptively zooms out to ensure complete segmentation of large structures. By detecting border changes and dynamically querying additional regions, it preserves global context while refining local details, mitigating the constraints of patch-wise processing.

and then refining predictions via a sliding window at full resolution. We introduce a more adaptive auto zoom strategy (Fig. 4). nnInteractive dynamically expands the region of interest (ROI) based on prediction borders, iteratively zooming out by a factor of 1.5 until the object is fully captured (up to 4 \times zoom out). The predicted low resolution mask is then resized to the original image resolution and refined with a sliding window approach. To optimize refinement, we process patches from most to least informative as measured by the amount of predicted foreground pixels they contain, ensuring the model starts in well-informed regions and incrementally extends to less constrained areas. The initial zoom factor is determined by the prompt that triggered it, ensuring the ROI encompasses the whole prompt plus a border of 1/6th of the patch size. This adaptive scheme minimizes computational overhead—small objects remain unaffected for faster inference, while large structures like the liver undergo progressive zoom and refinement. Unlike SegVol, nnInteractive dynamically adjusts the zoom level, preserving details, avoiding excessive down-sampling for small objects, and maintaining compatibility with large images without exceeding VRAM limits.

2.6. Ambiguity

Interactive segmentation models must adapt to user inputs and segment structures based on user intent, which is often ambiguous. For example, in liver tumor segmentation, users may segment the liver with or without the tumor, or the tumor alone. Similarly, in cardiac cine-MRI, the left ventricle may be segmented with or without its lumen. Models trained with rigid predefined class definitions struggle with such ambiguities (see Fig. A1). To address this, nnInteractive is trained with randomly sampled label variations, exposing the model to realistic anatomical combinations. This allows it to resolve ambiguities based on user interactions and flexibly adapt to different segmentation needs. Additionally, we do not harmonize class definitions across datasets, preserving ambiguities arising from differing annotation conventions and label definitions.

3. Training Data

3.1. Large-Scale, Multi-Domain Training Dataset

We develop our model on an unprecedented collection of over 120 publicly available 3D segmentation datasets, comprising a total of 64,518 volumes with 717,148 objects (5% of images held out for internal validation). This diverse collection spans a wide range of structures across multiple imaging modalities (Tab. A1). Specifically, the collection spans Computed Tomography (CT) [2, 5–7, 17–19, 35, 37, 41, 43, 45, 50–53, 56, 61, 66, 67, 69, 72, 74, 81, 83, 84, 86, 87, 89, 103, 105, 107, 108, 111, 113, 114, 117–120, 123, 128, 141, 143, 146–148, 151], different Magnetic Resonance Imaging (MRI) sequences [3, 6, 8–10, 15, 16, 21–23, 26, 33, 40, 50, 54, 55, 58, 63–65, 75, 77, 78, 90–92, 95, 96, 101, 102, 109, 124, 129, 130, 136–138, 145], 3D Ultrasound [12, 14, 31, 62, 68], Positron Emission Tomography (PET) [4, 35, 115] and 3D Microscopy [13, 36, 76, 80, 82, 94, 131, 132, 142, 149, 153]. The scale and diversity of this dataset collection is unmatched in 3D medical image segmentation, providing our model with exposure to a broad range of imaging modalities, (anatomical) structures, and pathological conditions. This enables robust, generalizable representations across varying image scales and clinical scenarios, laying the foundation for a versatile model capable of addressing diverse applications.

3.2. Label Diversity with SuperVoxels

Despite the scale and variety of our training distribution, further increasing robustness to unseen structures remains essential. To achieve this we incorporate pseudo-labels sampled with a probability of 0.2. While existing frameworks have relied on traditional computer vision algorithms such as Felzenszwalb [32] or SLIC [1], either applied directly to images [30, 144] or to encoded representations (Vista3D [42]), we leverage the capabilities of modern foundation models. Specifically, we utilize SAM’s automatic “segment everything” feature to generate high-confidence SuperVoxels ($\geq 92\%$) on axially sampled slices. We then employ SAM2’s video mask propagation, treating the remaining slices as sequential frames to generate a 3D segmentation. As illustrated in Fig. 5, our approach mitigates common limitations of SLIC, which struggles with image parcellation even when applied to embeddings, and Felzenszwalb, which tends to produce fuzzy borders. In contrast, our method generates high-quality, variable-sized objects, enhancing segmentation accuracy and adaptability across diverse structures.

4. Experiments

To evaluate nnInteractive we perform a large-scale comparison against established 2D and 3D models (Sec 4.1), benchmark on human expert scribbles (Sec 4.2), compare the per-

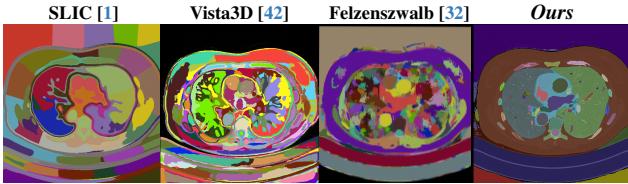


Figure 5. **SuperVoxels enhancing training label diversity.** Classical algorithms like SLIC and Felzenszwalb produce parcellation or fuzzy boundaries, even using image embeddings (Vista3D), while our approach yields precise variable-sized objects.

formance of supported prompting styles (Sec 4.3) and perform a User Study involving medical doctors (Sec 4.4).

4.1. Benchmarking against Established Methods.

For our primary comparison, we utilize the RadioActive Benchmark [135], which provides a comprehensive and reproducible evaluation framework with advanced prompting techniques, ensuring a fair and thorough comparison against existing methods. Competing 2D models include SAM [60] and SAM2 [112], originally trained on natural images, as well as medical adaptations: SAM-Med2D [24], trained on diverse biomedical modalities; MedSAM [88], trained with box prompts on 1.5M medical segmentations; and ScribblePrompt [144], trained on 65 medical datasets spanning healthy and pathological structures. For 3D models, we compare against SAM-Med3D [140], a 3D extension of SAM, and SegVol [30], trained on multi-organ and lesion datasets. Other notable interactive models, such as Vista3D [42], 3D SAM Adapter [38], and Prism [70], are designed for closed-set segmentation and were therefore not included in our evaluation. Our comparison is conducted on the ten held-out test datasets proposed by the benchmark [29, 79, 85, 97–99, 106, 126, 139, 152], covering standard CT and MRI modalities, including various anatomical structures and pathologies such as lesions. To further assess zero-shot adaptability and the generalization capabilities expected of foundation models, we extend the test set with out-of-distribution datasets featuring different imaging modalities and unseen tasks. Specifically, we evaluate models on microCT images of mouse tumors [49], knee MRI ligament structures [28] (unseen during training), microCT scans of insect anatomy [133], and PET images of head & neck tumors [59] (Tab A2). These datasets introduce significant domain shifts in resolution, contrast, target, and anatomical scale, providing a challenging benchmark for evaluating model robustness beyond conventional medical imaging tasks.

We start by comparing static point and 3D box prompts, as these are the most widely supported prompt types in current methods (e.g., MedSam [88] is limited to non-interactive boxes). In terms of points, 3D models receive a single point

per structure of interest, while 2D models receive one point per slice, which increases theoretically invested user effort. For bounding box interactions, each model receives the precise 3D bounding box around the target structure, while 2D models are queried with the corresponding sliced bounding box. Since 3D bounding boxes are impractical for interactive segmentation compromising user-friendliness, we trained a dedicated nnInteractive variant solely for benchmarking against existing models that rely on them (see Appendix A2). Following the benchmark’s guidelines, we also simulate static scribbles as out-of-plane lines/curves, querying the respective point in each slice for 2D models, while 3D models receive the full scribble.

To benchmark interactive refinement capabilities we use point prompts as they are the only prompting style supported across all methods. We place an initial point, followed up by 5 additional clicks sampled according to the model’s generated segmentation mask. Negative points are enabled if supported by the model.

4.2. Expert Scribble Benchmark.

To assess performance on real rather than simulated scribbles, we evaluate on unseen data from the MS-CMRSeg 2019 Challenge [154] with expert-provided per-slice axial scribble annotations for the left ventricle (LV), right ventricle (RV), and myocardium (MYO) [150]. ScribblePrompt [144], the only competing model with native scribble support, is prompted in a slice-wise 2D manner, using all scribbled slices. nnInteractive is evaluated using all annotated slices as well, but also tested with *only three scribbles* (top, middle, bottom), leveraging its ability to handle sparse prompts, simulating a *significantly reduced annotation effort*.

4.3. Evaluation of Prompting Styles

Current state-of-the-art methods are limited in their interactions, making it difficult to directly compare all interactions supported in nnInteractive for which there is no analogue in existing methods. This does not only affect the newly proposed lasso interaction but also the 2D bounding boxes and scribbles where nnInteractive is currently the only method able to accept these in a true 3D setting. We evaluate the usefulness of these advanced interactions on the aforementioned test and OOD datasets. We furthermore test random combination of interactions given to nnInteractive to confirm that it can make effective use of arbitrary combinations of inputs. Point, 2D bbox, lasso and scribble interactions are simulated using the logic outlined in Sec 2.3.

4.4. Radiological User Study.

To evaluate real-world usability, we conducted a user study on the segmentation of 12 tumor lesions across various anatomical regions, derived from MR and CT scans used in

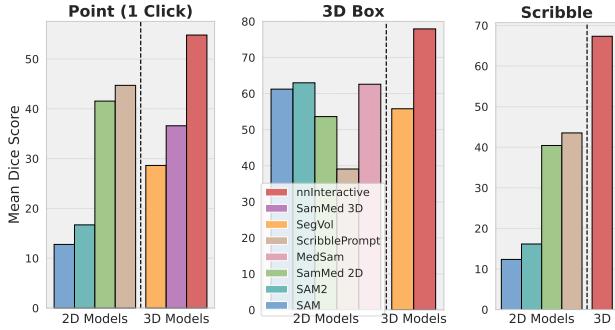


Figure 6. **Single-prompt performance.** Mean Dice scores on unseen test data when prompting with (1) a single point (2D models receive one per slice, giving them a theoretical advantage over 3D models), (2) a 3D bounding box (sliced for 2D models), and (3) an out-of-plane scribble (interpreted as points per slice in 2D). nnInteractive consistently outperforms all baselines, demonstrating superior segmentation quality across all interactions.

radiation therapy planning. After briefly introducing the radiologist to the prompt types, each lesion was annotated using nnInteractive and compared against expert manual segmentations from two raters —a resident and a specialist radiation oncologist —using their inter-rater variability as a reference. Additionally, we measured the time required for both manual and prompt-guided segmentation to assess efficiency improvements.

5. Results & Discussion

We first present results from the expanded RadioActive [135] benchmark and expert scribble comparison, followed by an analysis of nnInteractive’s unique contributions and our user study.

5.1. Comparison with SOTA

Static prompts. Figure 6 compares model performance using a single point, box, or out-of-plane scribble. nnInteractive consistently outperforms all baselines across interaction types. For point prompts, it surpasses the closest competitor, ScribblePrompt, by 10.1 Dice points, despite 2D models receiving significantly more prompts. For boxes, it achieves a 14.9 Dice point advantage over the second-best model SAM2, and for scribbles, an impressive 23.8-point lead. Notably, no single competing model consistently ranks second, as their performance highly varies by interaction type.

Interactive Prompts. Fig. 8 (Left) evaluates the interactive refinement capabilities of models using positive and negative point prompts, if supported. nnInteractive consistently achieves the highest performance, starting with a superior initial Dice score and reaching a Dice of more than 70 across all datasets. This surpasses the best-performing 2D model, ScribblePrompt, by 11.2 Dice points despite its sig-

| Prompt | Model | LV | RV | MYO | Avg. |
|------------|----------------|--------------|--------------|--------------|--------------|
| All Slices | ScribblePrompt | 66.08 | 90.04 | 86.82 | 80.98 |
| All Slices | nnInteractive | 78.86 | 92.93 | 90.07 | 87.29 |
| 3 Slices | nnInteractive | 74.40 | 91.24 | 87.33 | 84.29 |

Table 1. **Expert Scribbles Benchmark.** Performance (Dice) using human axial scribbles as prompts for the left & right ventricle (LV, RV) and myocardium (MYO). nnInteractive outperforms ScribblePrompt across all structures, even when limited to just three annotated slices instead of all scribbles.

nificantly higher number of prompts, and outperforms the strongest 3D competitor, SegVol, by 14.9 Dice points.

Expert Scribble Benchmark. Table 1 compares nnInteractive and ScribblePrompt on expert-drawn scribbles across three target classes. When provided with the same number of prompts (scribbles on every slice), nnInteractive outperforms ScribblePrompt by nearly 7 Dice points on average. Notably, even with only three annotated slices, nnInteractive achieves a Dice score of 84.3—surpassing ScribblePrompt by 3 points while requiring significantly less user interaction. This demonstrates nnInteractive’s ability to *achieve high segmentation accuracy with minimal annotation effort*.

5.2. Evaluation nnInteractive’s contributions

Prompting Styles. We evaluate nnInteractive’s prompting styles on the Test and OOD datasets using simulated interactions as described in 2.3. Due to the absence of competing methods capable of processing 2D bounding boxes, scribbles, and lasso in a 3D setting, we ground results using the only common interaction type: points. As shown in Fig. 8 (Right), among all interaction types, lasso performs best, achieving the highest Dice scores across all iterations with an AUC of 83.42. 2D bounding boxes yield high initial Dice scores but fall behind with more iterations, likely due to their coarse guidance being less effective for refinement. Scribbles initially provide less information than lasso or bounding boxes but surpass bounding boxes after five iterations, demonstrating their strength in precise refinement. Points perform the weakest, with an AUC of 71.76, consistently lagging behind other interaction types. This finding is notable as nnInteractive substantially outperforms current state-of-the-art models in points (both in static and interactive setting), despite points being its worst performing prompting style. Finally, we tested randomly selecting interactions each iteration, showing that interactions can be freely mixed in nnInteractive, with Rand(Lasso, Scribble, 2D Bbox) achieving Dice scores between the respective non-random interaction simulations.

AutoZoom. As shown in Fig. A3, AutoZoom with refinement effectively captures large objects and accelerates

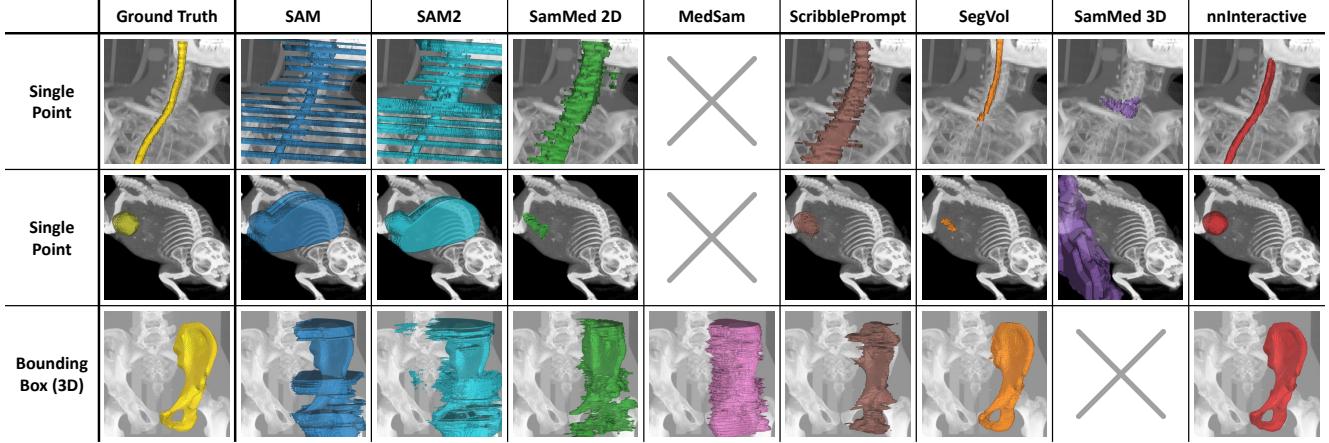


Figure 7. Qualitative comparison of interactive segmentation methods on unseen test images with different static prompting strategies: points and 3D boxes. nnInteractive achieves the highest accuracy, closely matching the ground truth, while others struggle with precision, consistency, or volumetric adaptation. Omitted results (\times) indicate unsupported prompt types.

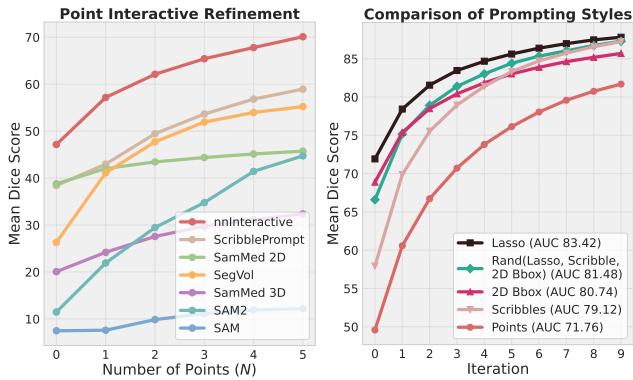


Figure 8. **Interactive Performance.** nnInteractive achieves the highest Dice scores in point-based refinement (Left), with a large gap over all competitors, despite 2D models receiving N points per slice and points being nnInteractive’s weakest prompt type (Right). Beyond points, nnInteractive excels with stronger interactions like scribbles and lasso. Interactions can be freely mixed for flexibility.

convergence with fewer user interactions. While AUC improvement on the Test and OOD datasets is minimal (82.72 vs 82.36) due to the predominance of small objects, substantial gains are observed for datasets with larger objects, such as HCC Tace liver [97] (AUC 95.40 vs 91.92) and InsectAnatomy [133] (AUC 94.81 vs 92.54). Further information in Appendix A6.1.

Resolving Ambiguity. Existing methods struggle with ambiguities due to overfitting on learned classes, for example failing to segment structures like kidneys and kidney tumors simultaneously (see Fig. A1). nnInteractive overcomes this limitation and can dynamically adapt to user input, efficiently resolving ambiguities with minimal iterations. We demonstrate this on two organs with and without

tumors in previously unseen images (Fig. A2).

Inference Time. Optimized for broad adoption, our implementation maintains VRAM usage below 10 GB (< 6 GB for small objects). Small structures like tumors and organs take merely 120–200 ms on an NVIDIA RTX 4090. For larger objects, AutoZoom iteratively zooms out and then refines segmentations, increasing inference time up to 1160 ms for a liver CT and 3700 ms in rare high-resolution cases. See Appendix A5.

5.3. Real-World Impact on Radiological Tasks

To assess the clinical applicability of nnInteractive, we compared its segmentation accuracy and efficiency against expert manual tumor annotation. Using Dice similarity coefficients, we found that segmentations generated using nnInteractive were as consistent with specialist annotations as inter-expert agreement itself. Median Dice scores were 0.842 ± 0.058 between resident and specialist annotations, 0.794 ± 0.040 between specialist annotations and nnInteractive, and 0.853 ± 0.068 between the resident and nnInteractive. Wilcoxon tests confirmed no significant differences between these comparisons ($p = 0.577$ and $p = 0.365$), indicating that *nnInteractive achieves expert-level performance*. Beyond accuracy, it dramatically improves efficiency: Experts completed per case segmentations in 179 ± 114 seconds using nnInteractive—72% faster than the 635 ± 343 seconds required for manual annotation. This substantial time reduction demonstrates nnInteractive’s potential to streamline clinical workflows, reducing workload while maintaining expert-grade precision.

6. Conclusion

We introduced nnInteractive, a universal 3D promptable segmentation framework that sets a new standard for AI-

driven interactive segmentation. By supporting diverse prompt types—including points, scribbles, bounding boxes, and lasso—nnInteractive bridges the gap between intuitive 2D interactions and full 3D volumetric segmentation. Trained on an unprecedented dataset of 120+ multimodal 3D datasets, our model achieves superior performance and usability across a wide range of imaging tasks. Extensive benchmarking demonstrates that our approach far surpasses existing methods, offering state-of-the-art segmentation accuracy while significantly reducing annotation effort. Integrated into established imaging platforms such as Nipari and MITK Workbench, it ensures seamless real-world adoption in clinical and research workflows, paving the way for more efficient and accessible 3D segmentation.

References

- [1] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012. [5](#) [6](#)
- [2] Hugo J. W. L. Aerts, Leonard Wee, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, Johan Bussink, Ren Monshouwer, Benjamin Haibe-Kains, Derek Rietveld, Frank Hoebers, Michelle M. Rietbergen, C. Ren Leemans, Andre Dekker, John Quackenbush, Robert J. Gillies, and Philippe Lambin. Data from nsclc-radiomics, 2019. [5](#) [22](#)
- [3] Bonnie Alexander, Wai Yen Loh, Lillian G Matthews, Andrea L Murray, Chris Adamson, Richard Beare, Jian Chen, Claire E Kelly, Peter J Anderson, Lex W Doyle, et al. Desikan-killiany-tourville atlas compatible version of m-crib neonatal parcellated whole brain atlas: The m-crib 2.0. *Frontiers in Neuroscience*, 13:34, 2019. [5](#) [22](#)
- [4] Vincent Andrarczyk, Valentin Oreiller, Moamen Abobakr, Azadeh Akhavanallaf, Panagiotis Balermpas, Sarah Boughdad, Leo Capriotti, Joel Castelli, Catherine Cheze Le Rest, and Pierre et al. Decazes. Overview of the HECKTOR challenge at MICCAI 2022: Automatic head and neck TumOR segmentation and outcome prediction in PET/CT. *Head Neck Tumor Chall* (2022), 2023. [5](#) [22](#)
- [5] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, and et al. The medical segmentation decathlon. *arXiv:2106.05735*, 2021. [5](#) [18](#) [22](#)
- [6] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M. Summers, and et al. The medical segmentation decathlon. *Nature Communications*, 2022. [5](#)
- [7] Armato III, Samuel G., Geoffrey McLennan, Luc Bidaut, Michael F. McNitt-Gray, Charles R. Meyer, Anthony P. Reeves, Binsheng Zhao, Denise R. Aberle, Claudia I. Henschke, and Eric A. et al. Hoffman. Data from lidec-idri, 2015. [5](#) [22](#)
- [8] Ujjwal Baid, Satyam Ghodasara, Suyash Mohan, Michel Bilello, Evan Calabrese, Errol Colak, Keyvan Farahani, Jayashree Kalpathy-Cramer, Felipe C Kitamura, Sarthak Pati, and Spyridon et al. The RSNA-ASNR-MICCAI BraTS 2021 benchmark on brain tumor segmentation and radiogenomic classification. *arXiv*, 2021. [5](#) [22](#)
- [9] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycski, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data*, 2017. [22](#)
- [10] Spyridon Bakas, Chiharu Sako, Hamed Akbari, Michel Bilello, Aristeidis Sotiras, Gaurav Shukla, Jeffrey D. Rudie, Natali Flores Santamaría, Anahita Fathi Kazerooni, Sarthak Pati, Saima Rathore, Elizabeth Mamourian, Sung Min Ha, William Parker, Jimit Doshi, Ujjwal Baid, Mark Bergman, Zev A. Binder, Ragini Verma, Robert A. Lustig, Arati S. Desai, Stephen J. Bagley, Zissimos Mourelatos, Jennifer Morrisette, Christopher D. Watt, Steven Brem, Ronald L. Wolf, Elias R. Melhem, MacLean P. Nasrallah, Suyash Mohan, Donald M. O'Rourke, and Christos Davatzikos. The university of pennsylvania glioblastoma (upenn-gbm) cohort: advanced mri, clinical, genomics, & radiomics. *Scientific Data*, 9(1):453, 2022. [5](#) [22](#)
- [11] Pedro R. A. S. Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xi-aomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation?, 2024. [2](#)
- [12] Bahareh Behboodi, Francois-xavier Carton, Matthieu Chabanais, Sandrine de Ribaupierre, Ole Solheim, Bodil KR Munkvold, Hassan Rivaz, Yiming Xiao, and Ingerid Reinertsen. Open access segmentations of intraoperative brain tumor ultrasound images. *Medical Physics*, 51(9):6525–6532, 2024. [5](#) [22](#)
- [13] Reinhard R Beichel, Robb W Glenny, Christian Bauer, and Melissa A Krueger. Lung anatomy + particle deposition (lapd) mouse archive, 2019. [5](#) [22](#)
- [14] Olivier Bernard, Johan G Bosch, Brecht Heyde, Martino Alessandrini, Daniel Barbosa, Sorina Camarasu-Pop, Frederic Cervenansky, Sébastien Valette, Oana Mirea, Michel Bernier, et al. Standardized evaluation system for left ventricular segmentation algorithms in 3d echocardiography. *IEEE transactions on medical imaging*, 35(4):967–977, 2015. [5](#) [22](#)

- [15] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, and et al. Gonzalez Ballester. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 2018. 5, 22
- [16] N. Bloch, A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, and K. Farahani. Nci-isbi 2013 challenge: Automated segmentation of prostate structures, 2015. 5, 22
- [17] Federico Bolelli, Luca Lumetti, Shankeeth Vinayahalingam, Mattia Di Bartolomeo, Arrigo Pellacani, Kevin Marchesini, Niels Van Nistelrooij, Pieter Van Lierop, Tong Xi, Yusheng Liu, Rui Xin, Tao Yang, Lisheng Wang, Haoshen Wang, Chenfan Xu, Zhiming Cui, Marek Wodzinski, Henning Müller, Yannick Kirchhoff, Maximilian R. Rokuss, Klaus Maier-Hein, Jaehwan Han, Wan Kim, Hong-Gi Ahn, Tomasz Szczepański, Michal K. Grzeszczyk, Przemyslaw Korzeniowski, Vicent Caselles Ballester, Xavier Paolo Burgos-Artizzu, Ferran Prados Carrasco, Stefaan Berge, Bram Van Ginneken, Alexandre Anesi, and Costantino Grana. Segmenting the inferior alveolar canal in cbct volumes: the toothfairy challenge. *IEEE Transactions on Medical Imaging*, pages 1–1, 2024. 5, 22
- [18] David Bouget, Arve Jørgensen, Gabriel Kiss, Haakon Olav Leira, and Thomas Langø. Semantic segmentation and detection of mediastinal lymph nodes and anatomical structures in ct data for lung cancer staging. *International journal of computer assisted radiology and surgery*, 14:977–986, 2019. 22
- [19] David Bouget, André Pedersen, Johanna Vanel, Haakon O. Leira, and Thomas Langø. Mediastinal lymph nodes segmentation using 3d convolutional neural network ensembles and anatomical priors guiding. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 0(0):1–15, 2022. 5, 22
- [20] Clint A Boyd. The cranial anatomy of the neornithischian dinosaur theselosaurus neglectus. *PeerJ*, 2:e669, 2014. 26
- [21] Mateusz Buda, Ashirbani Saha, and Maciej A Mazurowski. Association of genomic subtypes of lower-grade gliomas with shape features automatically extracted by a deep learning algorithm. *Computers in biology and medicine*, 109: 218–225, 2019. 5, 22
- [22] Victor M. Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M. Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, and et.al Ma. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The mms challenge. *IEEE Transactions on Medical Imaging*, 2021. 22
- [23] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L. Cuzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, and et al. Sudre. Longitudinal multiple sclerosis lesion segmentation: Resource and challenge. *NeuroImage*, 2017. 5, 22
- [24] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Ji long Chen, Lei Jiang, Hui Sun, Junjun He, Shaoting Zhang, Min Zhu, and Yu Qiao. Sam-med2d, 2023. 2, 6
- [25] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository. *Journal of Digital Imaging*, 2013. 22
- [26] Tugba Akinci D’Antonoli, Lucas K. Berger, Ashraya K. Indrakanti, Nathan Vishwanathan, Jakob Weiß, Matthias Jung, Zeynep Berkarda, Alexander Rau, Marco Reisert, Thomas Küstner, Alexandra Walter, Elmar M. Merkle, Martin Seegeroth, Joshy Cyriac, Shan Yang, and Jakob Wasserthal. Totalsegmentator mri: Sequence-independent segmentation of 59 anatomical structures in mr images, 2024. 5, 22
- [27] MJJ de Grauw, E Th Scholten, EJ Smit, MJCM Rutten, M Prokop, B van Ginneken, and A Hering. The uls23 challenge: a baseline model and benchmark dataset for 3d universal lesion segmentation in computed tomography. *arXiv preprint arXiv:2406.05231*, 2024. 2
- [28] Arjun D Desai, Andrew M Schmidt, Elka B Rubin, Christopher M Sandino, Marianne S Black, Valentina Mazzoli, Kathryn J Stevens, Robert Boutin, Christopher Ré, Garry E Gold, Brian A Hargreaves, and Akshay S Chaudhari. Skmtea: A dataset for accelerated mri reconstruction with dense image labels for quantitative clinical evaluation, 2022. 6, 23
- [29] Reuben Dorent, Roya Khajavi, Tagwa Idris, Erik Ziegler, Bhanusupriya Somarouthu, Heather Jacene, Ann LaCasce, Jonathan Deissler, Jan Ehrhardt, Sofija Engelson, Stefan M. Fischer, Yun Gu, Heinz Handels, Satoshi Kasai, Satoshi Kondo, Klaus Maier-Hein, Julia A. Schnabel, Guotai Wang, Litingyu Wang, Tassilo Wald, Guang-Zhong Yang, Hanxiao Zhang, Minghui Zhang, Steve Pieper, Gordon Harris, Ron Kikinis, and Tina Kapur. Lnq 2023 challenge: Benchmark of weakly-supervised techniques for mediastinal lymph node quantification, 2024. 2, 6, 23
- [30] Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. *arXiv preprint arXiv:2311.13385*, 2023. 2, 4, 5, 6
- [31] Vanessa Gonzalez Duque, Alexandra Marquardt, Jordanka Velikova, Lilian Lacourpaille, Antoine Nordez, Marion Crouzier, Hong Joo Lee, Diana Mateus, and Nassir Navab. Ultrasound segmentation analysis via distinct and completed anatomical borders. *International Journal of Computer Assisted Radiology and Surgery*, 19(7):1419–1427, 2024. 5, 22
- [32] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59:167–181, 2004. 5, 6
- [33] Lidia Garrucho, Claire-Anne Reidel, Kaisar Kushibar, Smriti Joshi, Richard Osuala, Apostolia Tsirikoglou, Maciej Bobowicz, Javier del Riego, Alessandro Catanese, Katarzyna Gwoździewicz, et al. Mama-mia: A large-scale multi-center breast cancer dce-mri benchmark dataset with

- expert segmentations. *arXiv preprint arXiv:2406.13844*, 2024. 5, 22
- [34] Sergios Gatidis and Thomas Kuestner. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions (FDG-PET-CT-Lesions), 2022. 22
- [35] andKuestner T.”Gatidis S. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions (fdg-pet-ct-lesions). The Cancer Imaging Archive,, 2022. 2, 5
- [36] Stephan Gerhard, Jan Funke, Julien Martel, Albert Cardona, and Richard Fetter. Segmented anisotropic sstem dataset of neural tissue. *figshare*, pages 0–0, 2013. 5, 22
- [37] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P. Pereira, Matthew J. Clarkson, and Dean C. Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE Transactions on Medical Imaging*, 2018. 5, 22
- [38] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. 3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation. *Medical Image Analysis*, 98:103324, 2024. 2, 6
- [39] Karol Gotkowski, Shuvam Gupta, Jose RA Godinho, Camila GS Toch trop, Klaus H Maier-Hein, and Fabian Isensee. Particleseg3d: a scalable out-of-the-box deep learning segmentation solution for individual particle characterization from micro ct images in mineral processing and recycling. *Powder Technology*, 434:119286, 2024. 22
- [40] Endre Grøvik, Darvin Yi, Michael Iv, Elizabeth Tong, Daniel Rubin, and Greg Zaharchuk. Deep learning enables automatic detection and segmentation of brain metastases on multisequence MRI. *J. Magn. Reson. Imaging*, 2020. 5, 22
- [41] Yuting He, Guanyu Yang, Jian Yang, Rongjun Ge, Youyoung Kong, Xiaomei Zhu, Shaobo Zhang, Pengfei Shao, Huazhong Shu, Jean-Louis Dillenseger, et al. Meta grayscale adaptive network for 3d integrated renal structures segmentation. *Medical image analysis*, 71:102055, 2021. 5, 22
- [42] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, Daguang Xu, and Wenqi Li. Vista3d: Versatile imaging segmentation and annotation model for 3d computed tomography, 2024. 2, 4, 5, 6, 19
- [43] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, and et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023. 5, 22
- [44] Ziyuan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaotong Zhang, and Yu Qiao. Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training, 2023. 1
- [45] Muhammad Imran, Jonathan R. Krebs, Veera Rajasekhar Reddy Gopu, Brian Fazzone, Vishal Balaji Sivarajan, Amarjeet Kumar, Chelsea Viscardi, Robert Evans Heithaus, Benjamin Shickel, Yuyin Zhou, Michol A. Cooper, and Wei Shao. Cis-unet: Multi-class segmentation of the aorta in computed tomography angiography via context-aware shifted window self-attention. *Computerized Medical Imaging and Graphics*, 118:102470, 2024. 5, 22
- [46] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2)(2):203–211, 2021. 1, 2, 17
- [47] Fabian Isensee, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus Maier-Hein, and Paul F. Jaeger. nnU-net revisited: A call for rigorous validation in 3d medical image segmentation, 2024. 1, 2, 17
- [48] Alexander Jaus, Constantin Seibold, Kelsey Hermann, Alexandra Walter, Kristina Giske, Johannes Haubold, Jens Kleesiek, and Rainer Stiefelhagen. Towards unifying anatomy segmentation: Automated generation of a full-body ct dataset via knowledge aggregation and anatomical guidelines. *arXiv preprint arXiv:2307.13375*, 2023. 22
- [49] Malte Jensen, Andreas Clemmensen, Jacob Gorm Hansen, Julie van Krimpen Mortensen, Emil N. Christensen, Andreas Kjaer, and Rasmus Sejersten Ripa. 3d whole body preclinical micro-ct database of subcutaneous tumors in mice with annotations from 3 annotators. *Scientific Data*, 11(1):1021, 2024. 6, 23
- [50] Yuanfeng Ji, Haotian Bai, Chongjian GE, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Advances in Neural Information Processing Systems*, 2022. 5, 22
- [51] Yuan Jin, Antonio Pepe, Jianning Li, Christina Gsaxner, Fen-Hua Zhao, Kelsey L Pomykala, Jens Kleesiek, Alejandro F Frangi, and Jan Egger. AI-based aortic vessel tree segmentation for cardiovascular diseases treatment: Status quo. *arXiv*, 2021. 22
- [52] Petr Jordan, Philip M Adamson, Vrunda Bhattbhatt, Surabhi Beriwal, Sangyu Shen, Oskar Radermecker, Supratik Bose, Linda S Strain, Michael Offe, David Fraley, et al. Pediatric chest-abdomen-pelvis and abdomen-pelvis ct images with expert organ contours. *Medical physics*, 49(5):3523–3528, 2022. 22
- [53] Ma Jun, Ge Cheng, Wang Yixin, An Xingle, Gao Jiantao, Yu Ziqi, Zhang Mingqing, Liu Xin, Deng Xueyuan, Cao Shucheng, Wei Hao, Mei Sen, Yang Xiaoyu, Nie Ziwei, Li Chen, Tian Lu, Zhu Yuntao, Zhu Qiongjie, Dong Guoqiang, and He Jian. COVID-19 CT Lung and Infection Segmentation Dataset, 2020. 5, 22
- [54] Parikshit Juvekar, Reuben Dorent, Fryderyk Kögl, Erickson Torio, Colton Barr, Laura Rigolo, Colin Galvin, Nick Jowkar, Anees Kazi, Nazim Haouchine, Harneet Cheema, Nassir Navab, Steve Pieper, William M. Wells, Wenyu Linda Bi, Alexandra Golby, Sarah Friskin, and Tina Kapur. Remind: The brain resection multimodal imaging database. *Scientific Data*, 11(1):494, 2024. 5, 22

- [55] Alan H Kadish, David Bello, J Paul Finn, Robert O Bonow, Andi Schaechter, Haris Subacius, Christine Albert, James P Daubert, Carissa G Fonseca, and Jeffrey J Goldberger. Rationale and design for the defibrillators to reduce risk by magnetic resonance imaging evaluation (determine) trial. *Journal of cardiovascular electrophysiology*, 20(9):982–987, 2009. 5, 18, 22
- [56] Aasheesh Kanwar, Brandon Merz, Cheryl Clauch, Shushan Rana, Arthur Hung, and Reid F Thompson. Stress-testing pelvic autosegmentation algorithms using anatomical edge cases. *Physics and Imaging in Radiation Oncology*, 25:100413, 2023. 5, 22
- [57] Alexandros Karargyris, Renato Umeton, Micah J. Sheller, Alejandro Aristizabal, Johnu George, Anna Wuest, Sarthak Pati, Hasan Kassem, Maximilian Zenk, Ujjwal Baid, and et al. Federated benchmarking of medical artificial intelligence with medperf. *Nature Machine Intelligence*, 2023. 22
- [58] Ali Emre Kavur, M. Alper Selver, Oguz Dicle, Mustafa Baris, and N. Sinem Gezer. Chaos - combined (ct-mr) healthy abdominal organ segmentation challenge data, 2019. 5, 22
- [59] P. Kinahan, M. Muzi, B. Bialecki, and L. Coombs. Data from the acrin 6685 trial hnsc-fdg-pet/ct. Data set, 2019. 6, 23
- [60] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2, 6
- [61] Kendall J Kiser, Sara Ahmed, Sonja Stieb, Abdallah SR Mohamed, Hesham Elhalawani, Peter YS Park, Nathan S Doyle, Brandon J Wang, Arko Barman, Zhao Li, et al. Plethora: Pleural effusion and thoracic cavity segmentations in diseased lungs for benchmarking chest ct processing pipelines. *Medical physics*, 47(11):5941–5952, 2020. 5, 22
- [62] Markus Krönke, Christine Eilers, Desislava Dimova, Melanie Köhler, Gabriel Buschner, Lilit Schweiger, Lemonia Konstantinidou, Marcus Makowski, James Nagarajah, Nassir Navab, et al. Tracked 3d ultrasound and deep neural network-based thyroid segmentation reduce interobserver variability in thyroid volumetry. *Plos one*, 17(7):e0268550, 2022. 5, 18, 22
- [63] HJ Kuijf, E Bennink, KL Vincken, N Weaver, GJ Biessels, and MA Viergever. Mr brain segmentation challenge 2018 data. DOI: <https://doi.org/10.34894/E0U32Q>, 2024. 5, 22
- [64] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. Standardized assessment of automatic segmentation of white matter hyperintensities and results of the wmh segmentation challenge. *IEEE transactions on medical imaging*, 38(11):2556–2568, 2019. 22
- [65] Alain Lalande, Zhihao Chen, Thomas Decourselle, Abdul Qayyum, Thibaut Pommier, Luc Lorgis, Ezequiel de La Rosa, Alexandre Cochet, Yves Cottin, Dominique Gin hac, et al. Emidec: a database usable for the automatic evaluation of myocardial infarction from delayed-enhancement cardiac mri. *Data*, 5(4):89, 2020. 5, 22
- [66] Z. Lambert, C. Petitjean, B. Dubray, and S. Ruan. Segthor: Segmentation of thoracic organs at risk in ct images. *arXiv:1912.05950*, 2019. 5, 22
- [67] Bennett Landman, Zhoubing Xu, Juan Eugenio Iglesias, Martin Styner, and et al. 2015 miccai multi-atlas labeling beyond the cranial vault workshop and challenge. In *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*, 2015. 5, 22
- [68] Sarah Leclerc, Erik Smistad, Joao Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, et al. Deep learning for segmentation using an open large-scale dataset in 2d echocardiography. *IEEE transactions on medical imaging*, 38(9):2198–2210, 2019. 5, 22
- [69] Hongsheng Li, Jinghao Zhou, Jincheng Deng, and Ming Chen. Automatic structure segmentation for radiotherapy planning challenge, 2019. <https://structseg2019.grand-challenge.org/> 25/02/2022). 5, 22
- [70] Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, and Ipek Oguz. Prism: A promptable and robust interactive segmentation model with visual prompts, 2024. 2, 6
- [71] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *The Twelfth International Conference on Learning Representations*, 2024. 22
- [72] Xiangyu Li, Gongning Luo, Kuanquan Wang, Hongyu Wang, Jun Liu, Xinjie Liang, Jie Jiang, Zhenghao Song, Chunyue Zheng, Haokai Chi, Mingwang Xu, Yingte He, Xinghua Ma, Jingwen Guo, Yifan Liu, Chuanpu Li, Zeli Chen, Md Mahfuzur Rahman Siddiquee, Andriy Myronenko, Antoine P. Sanner, Anirban Mukhopadhyay, Ahmed E. Othman, Xingyu Zhao, Weiping Liu, Jinhuang Zhang, Xiangyuan Ma, Qinghui Liu, Bradley J. MacIntosh, Wei Liang, Moona Mazher, Abdul Qayyum, Valeria Abramova, Xavier Lladó, and Shuo Li. The state-of-the-art 3d anisotropic intracranial hemorrhage segmentation on non-contrast head ct: The instance challenge, 2023. 5, 22
- [73] Wenjun Liao, Xiangde Luo, Yuan He, Ye Dong, Churong Li, Kang Li, Shichuan Zhang, Shaoting Zhang, Guotai Wang, and Jianghong Xiao. Comprehensive evaluation of a deep learning model for automatic organs-at-risk segmentation on heterogeneous computed tomography images for abdominal radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 117(4):994–1006, 2023. 22
- [74] Hans Liebl, David Schinz, Anjany Sekuboyina, Luca Malagutti, Maximilian T. Löffler, Amirhossein Bayat, Malek El Husseini, Giles Tetteh, Katharina Grau, and et al. Niederreiter. A computed tomography vertebral segmentation dataset with anatomical variations and multi-vendor scanner data, 2021. 5, 22
- [75] Sook-Lei Liew, Julia M Anglin, Nick W Banks, Matt Sondag, Kaori L Ito, Hosung Kim, Jennifer Chan, Joyce

- Ito, Connie Jung, Nima Khoshab, and at al. A large, open source dataset of stroke anatomical brain images and manual lesion segmentations. *Sci. Data*, 2018. [5, 22](#)
- [76] Zudi Lin, Donglai Wei, Mariela D. Petkova, Yuelong Wu, Zergham Ahmed, Krishna Swaroop K, Silin Zou, Nils Wendt, Jonathan Boulanger-Weill, Xueying Wang, Nagaraju Dhanyasi, Ignacio Arganda-Carreras, Florian Engert, Jeff Lichtman, and Hanspeter Pfister. *NucMM Dataset: 3D Neuronal Nuclei Instance Segmentation at Sub-Cubic Millimeter Scale*, page 164–174. Springer International Publishing, 2021. [5, 22](#)
- [77] Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, and et al. Zhang. Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. *Medical Image Analysis*, 2014. [5, 22](#)
- [78] Geert Litjens, Oscar Debats, Jelle Barentsz, Nico Karssemeijer, and Henkjan Huisman. Spie-aapm prostatex challenge data, 2017. [5, 22](#)
- [79] Yanzhen Liu, Sutuke Yibulayimu, Yudi Sang, Gang Zhu, Yu Wang, Chunpeng Zhao, and Xinbao Wu. Pelvic fracture segmentation using a multi-scale distance-weighted neural network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 312–321. Springer, 2023. [6, 23](#)
- [80] Vebjorn Ljosa, Katherine L Sokolnicki, and Anne E Carpenter. Annotated high-throughput microscopy image sets for validation. *Nature methods*, 9(7):637, 2012. [5, 18, 22](#)
- [81] Maximilian T. Löffler, Anjany Sekuboyina, Alina Jacob, Anna-Lena Grau, Andreas Scharr, Malek El Husseini, Mareike Kallweit, Claus Zimmer, Thomas Baum, and Jan S. Kirschke. A vertebral segmentation dataset with fracture grading. *Radiology: Artificial Intelligence*, 2020. [5, 22](#)
- [82] Aurélien Lucchi, Kevin Smith, Radhakrishna Achanta, Graham Knott, and Pascal Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE transactions on medical imaging*, 31(2):474–486, 2011. [5, 22](#)
- [83] Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, et al. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge. *arXiv preprint arXiv:2304.03708*, 2023. [5, 22](#)
- [84] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N. Metaxas, Guotai Wang, and Shaoting Zhang. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *Medical Image Analysis*, 2022. [5, 22](#)
- [85] Xiangde Luo, Jia Fu, Yunxin Zhong, Shuolin Liu, Bing Han, Mehdi Astaraki, Simone Bendazzoli, Iuliana Toma-Dasu, Yiwen Ye, Ziyang Chen, et al. Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *arXiv preprint arXiv:2312.09576*, 2023. [6, 23](#)
- [86] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, Shucheng Cao, Qi Zhang, Shangqing Liu, Yunpeng Wang, Yuhui Li, Jian He, and Xiaoping Yang. Abdomenct-1k: Is abdominal organ segmentation a solved problem? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. [5, 22](#)
- [87] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, Fan Zhang, Wentao Liu, and YuanKe Pan ant et al. Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge. *arXiv preprint arXiv:2308.05862*, 2023. [5, 22](#)
- [88] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. [2, 6](#)
- [89] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Ershuai Wang, Qin Zhou, Ziyan Huang, Pengju Lyu, Jian He, and Bo Wang. Automatic organ and pan-cancer segmentation in abdomen ct: the flare 2023 challenge, 2024. [5](#)
- [90] Jacob A. Macdonald, Zhe Zhu, Brandon Konkel, Maciej Mazurowski, Walter Wiggins, and Mustafa Bashir. Duke liver dataset (mri) v2, 2020. [5, 22](#)
- [91] Oskar Maier, Matthias Wilms, Janina von der Gablentz, Ulrike M. Krämer, Thomas F. Münte, and Heinz Handels. Extra tree forests for sub-acute ischemic stroke lesion segmentation in mr sequences. *Journal of Neuroscience Methods*, 2015. [22](#)
- [92] Daniel S Marcus, Tracy H Wang, Jamie Parker, John G Csernansky, John C Morris, and Randy L Buckner. Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007. [5, 22](#)
- [93] Carlos Martín-Isla, Víctor M. Campello, Cristian Izquierdo, Kaisar Kushibar, Carla Sendra-Balcells, Polyxeni Gkontra, Alireza Sojoudi, Mitchell J. Fulton, Tewodros Weldebirhan Arega, and et al. Punithakumar. Deep learning segmentation of the right ventricle in cardiac mri: The mms challenge. *IEEE Journal of Biomedical and Health Informatics*, 2023. [22](#)
- [94] Martin Maška, Vladimír Ulman, Pablo Delgado-Rodriguez, Estibaliz Gómez-de Mariscal, Tereza Nečasová, Fidel A Guerrero Peña, Tsang Ing Ren, Elliot M Meyerowitz, Tim Scherr, Katharina Löffler, et al. The cell tracking challenge: 10 years of objective benchmarking. *Nature Methods*, 20(7):1010–1020, 2023. [5, 22](#)
- [95] N. Mayr, W. T. C. Yuh, S. Bowen, M. Harkenrider, M. V. Knopp, E. Y.-P. Lee, E. Leung, S. S. Lo, W. Small Jr., and A. H. Wolfson. Cervical cancer – tumor heterogeneity: Serial functional and molecular imaging across the radiation therapy course in advanced cervical cancer (version 1), 2023. Data set. [5, 22](#)
- [96] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, and et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging*, 2015. [5, 22](#)
- [97] Ahmed W. Moawad, David Fuentes, Ali Morshid, Ahmed M. Khalaf, Mohab M. Elmohr, Abdelrahman Abu-

- saif, John D. Hazle, Ahmed O. Kaseb, Manal Hassan, Armeen Mahvash, Janio Szklaruk, Aliyya Qayyom, and Khaled Elsayes. Multimodality annotated hcc cases with and without advanced imaging segmentation, 2021. 6, 8, 23
- [98] Ahmed W. Moawad, Ayahallah A. Ahmed, Mohab ElMohr, Mohamed Eltaher, Mohammed Amir Habra, Sarah Fisher, Nancy Perrier, Miao Zhang, David Fuentes, and Khaled Elsayes. Voxel-level segmentation of pathologically-proven adrenocortical carcinoma with ki-67 expression (adrenal-acc-ki67-seg), 2023. 23
- [99] Ali M. Muslim, Syamsiah Mashohor, Gheyath Al Gawaym, Rozi Mahmud, Marsyita binti Hanafi, Osama Al-nuaimi, Raad Josephine, and Abdullah Dhaifallah Almutairi. Brain mri dataset of multiple sclerosis with consensus manual lesion segmentation and patient meta information. *Data in Brief*, 2022. 6, 23
- [100] Marco Nolden, Sascha Zelzer, Alexander Seitel, Diana Wald, Michael Müller, Alfred M Franz, Daniel Maleike, Markus Fangerau, Matthias Baumhauer, Lena Maier-Hein, et al. The medical imaging interaction toolkit: challenges and advances: 10 years of open-source development. *International journal of computer assisted radiology and surgery*, 8:607–620, 2013. 2
- [101] Danielle F Pace, Hannah TM Contreras, Jennifer Romanowicz, Shruti Ghelani, Imon Rahaman, Yue Zhang, Patricia Gao, Mohammad Imrul Jubair, Tom Yeh, Polina Goldland, et al. Hvsmr-2.0: A 3d cardiovascular mr dataset for whole-heart segmentation in congenital heart disease. *Scientific Data*, 11(1):721, 2024. 5, 22
- [102] Kelly Payette, Priscille de Dumast, Hamza Kebiri, Ivan Ezhev, Johannes C Paetzold, Suprosanna Shit, Asim Iqbal, Romesa Khan, Raimund Kottke, Patrice Grehten, et al. An automatic multi-tissue human fetal brain segmentation benchmark using the fetal tissue annotation dataset. *Scientific data*, 8(1):167, 2021. 5, 22
- [103] Joao Pedrosa, Guilherme Aresta, Carlos Ferreira, Gurraj Atwal, Hady Ahmady Phoulady, Xiaoyu Chen, Rongzhen Chen, Jiaoliang Li, Liansheng Wang, Adrian Galdran, et al. Lndb challenge on automatic lung cancer patient management. *Medical image analysis*, 70:102027, 2021. 5, 22
- [104] Antonio Pepe, Jianning Li, Malte Rolf-Pissarczyk, Christina Gsaxner, Xiaojun Chen, Gerhard A Holzapfel, and Jan Egger. Detection, segmentation, simulation and visualization of aortic dissections: A review. *Med. Image Anal.*, 2020. 22
- [105] S. Pieper, N. Haouchine, D. B. Hackney, W. M. Wells, M. Sanhinova, T. Balboni, A. Spektor, M. Huynh, S. Tanguuri, E. Kim, J. P. Guenette, D. E. Kozono, B. Czajkowski, S. Caplan, P. Doyle, H. Kang, and R. N. Alkalay. Spine metastatic bone cancer: pre and post radiotherapy ct (spine-mets-ct-seg) [dataset] (version 1), 2024. 5, 22
- [106] Gašper Podobnik, Primož Strojan, Primož Peterlin, Bulat Ibragimov, and Tomaž Vrtovec. Han-seg: The head and neck organ-at-risk ct and mr segmentation dataset. *Medical physics*, 50(3):1917–1927, 2023. 6, 23
- [107] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks, 2023. 5
- [108] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 ct volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, 2023. 5, 22
- [109] Félix Quinton, Romain Popoff, Benoît Presles, Sarah Leclerc, Fabrice Meriaudeau, Guillaume Nodari, Olivier Lopez, Julie Pellegrinelli, Olivier Chevallier, Dominique Ginhac, et al. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5):79, 2023. 5, 22
- [110] Lukas Radl, Yuan Jin, Antonio Pepe, Jianning Li, Christina Gsaxner, Fen-Hua Zhao, and Jan Egger. AVT: Multicenter aortic vessel tree CTA dataset collection with ground truth segmentation masks. *Data Brief*, 2022. 22
- [111] Patrik F Raudaschl, Paolo Zaffino, Gregory C Sharp, Maria Francesca Spadea, Antong Chen, Benoit M Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, et al. Evaluation of segmentation methods on head and neck ct: auto-segmentation challenge 2015. *Medical physics*, 44(5):2020–2036, 2017. 5, 22
- [112] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos, 2024. 2, 6
- [113] M. Riera-Marín, J.-M. Kleiß, A. Aubanell, and A. Antolín. Curvas dataset (v1.0.1). MEDICAL IMAGE COMPUTING AND COMPUTER ASSISTED INTERVENTION (MICCAI), Marrakesch, 2024. Data set. 5, 22
- [114] Blaine Rister, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. CT-ORG: A dataset of CT volumes with multiple organ segmentations, 2019. 5, 22
- [115] Maximilian Rokuss, Balint Kovacs, Yannick Kirchhoff, Shuhan Xiao, Constantin Ulrich, Klaus H. Maier-Hein, and Fabian Isensee. From fdg to psma: A hitchhiker’s guide to multitracer, multicenter lesion segmentation in pet/ct imaging, 2024. 5
- [116] Maximilian Rokuss, Yannick Kirchhoff, Seval Akbal, Balint Kovacs, Saikat Roy, Constantin Ulrich, Tassilo Wald, Lukas T. Rotkopf, Heinz-Peter Schlemmer, and Klaus Maier-Hein. Lesionlocator: Zero-shot universal tumor segmentation and tracking in 3d whole-body imaging, 2025. 2
- [117] Holger R. Roth, Le Lu, Ari Seff, Kevin M. Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and Ronald M. Summers. *A New 2.5D Representation for Lymph Node Detection Using Random Sets of Deep Convolutional Neural Network Observations*, page 520–527. Springer International Publishing, 2014. 5, 22
- [118] Holger R Roth, Le Lu, Ari Seff, Kevin M Cherry, Joanne Hoffman, Shijun Wang, Jiamin Liu, Evrim Turkbey, and

- Ronald M Summers. A new 2.5 d representation for lymph node detection using random sets of deep convolutional neural network observations. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14–18, 2014, Proceedings, Part I* 17, pages 520–527. Springer, 2014. 22
- [119] Holger R. Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B. Turkbey, and Ronald M. Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, 2015. 22
- [120] Holger R. Roth, Ziyue Xu, Carlos Tor-Díez, Ramon Sanchez Jacob, Jonathan Zember, Jose Molto, Wenqi Li, Sheng Xu, Baris Turkbey, Evrim Turkbey, Dong Yang, Ahmed Harouni, Nicola Rieke, Shishuai Hu, Fabian Isensee, Claire Tang, Qinji Yu, Jan Sölder, Tong Zheng, Vitali Liauchuk, Ziqi Zhou, Jan Hendrik Moltz, Bruno Oliveira, Yong Xia, Klaus H. Maier-Hein, Qikai Li, Andreas Husch, Luyang Zhang, Vassili Kovalev, Li Kang, Alessa Hering, João L. Vilaça, Mona Flores, Daguang Xu, Bradford Wood, and Marius George Linguraru. Rapid artificial intelligence solutions in a pandemic—the covid-19-20 lung ct lesion segmentation challenge. *Medical Image Analysis*, 82:102605, 2022. 5, 22
- [121] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R. Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H. Maier-Hein. Sam.md: Zero-shot medical image segmentation capabilities of the segment anything model, 2023. 2
- [122] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, and Klaus Maier-Hein. Mednext: Transformer-driven scaling of convnets for medical image segmentation, 2024. 1
- [123] Anjany Sekuboyina, Malek E. Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, Martin Urschler, Maodong Chen, Dalong Cheng, and et al. Lessmann. Verse: A vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical Image Analysis*, 2021. 5, 22
- [124] Jonathan Shapey, Aaron Kujawa, Reuben Dorent, Guotai Wang, Sotirios Bisdas, Alexis Dimitriadis, Diana Grishchuck, Ian Paddick, Neil Kitchen, Robert Bradford, Shaikel Saeed, Sebastien Ourselin, and Tom Vercauteren. Segmentation of vestibular schwannoma from magnetic resonance imaging: An open annotated dataset and baseline algorithm (vestibular-schwannoma-SEG), 2021. 5, 22
- [125] Amber L. Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram van Ginneken, Annette Kopp-Schneider, Bennett A. Landman, Geert Litjens, Bjoern Menze, and et al. A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv:1902.09063*, 2019. 22
- [126] Amber L. Simpson, Jacob Peoples, John M. Creasy, Gabor Fichtinger, Natalie Gangai, Andras Lasso, Krishna Nand Keshava Murthy, Jinru Shia, Michael I. D'Angelica, and Richard K. G. Do. Preoperative ct and survival data for patients undergoing resection of colorectal liver metastases (colorectal-liver-metastases), 2023. 6, 23
- [127] Nicholas Sofroniew, Talley Lambert, Grzegorz Bokota, Juan Nunez-Iglesias, Peter Sobolewski, Andrew Sweet, Lorenzo Gaifas, Kira Evans, Alister Burt, Draga Doncila Pop, Kevin Yamauchi, Melissa Weber Mendonça, Genevieve Buckley, Wouter-Michiel Vierdag, Loic Royer, Ahmet Can Solak, Kyle I. S. Harrington, Jannis Ahlers, Daniel Althviz Moré, Oren Amsalem, Ashley Anderson, Andrew Annex, Peter Boone, Jordão Bragantini, Matthias Bussonnier, Clément Caporal, Jan Eglinger, Andreas Eisenbarth, Jeremy Freeman, Christoph Gohlke, Kabilan Gunalan, Hagai Har-Gil, Mark Harfouche, Volker Hilsenstein, Katherine Hutchings, Jessy Lauer, Gregor Lichtner, Ziyang Liu, Lucy Liu, Alan Lowe, Luca Marconato, Sean Martin, Abigail McGovern, Lukasz Migas, Nadalyn Miller, Hector Muñoz, Jan-Hendrik Müller, Christopher Nauroth-Kreß, David Palecek, Constantin Pape, Eric Perlman, Kim Pevey, Gonzalo Peña-Castellanos, Andrea Pierré, David Pinto, Jaime Rodríguez-Guerra, David Ross, Craig T. Russell, James Ryan, Gabriel Selzer, MB Smith, Paul Smith, Konstantin Sofiuk, Johannes Soltwedel, David Stansby, Jules Vanaret, Pam Wadhwa, Martin Weigert, Jonas Windhager, Philip Winston, and Rubin Zhao. napari: a multi-dimensional image viewer for python, 2025. 2
- [128] Karen-Helene Støverud, David Bouget, Andre Pedersen, Håkon Olav Leira, Thomas Langø, and Erlend Fagertun Hofstad. AeroPath: An airway segmentation benchmark dataset with challenging pathology, 2023. 5, 22
- [129] Carole H Sudre, Kimberlin Van Wijnen, Florian Dubost, Hieab Adams, David Atkinson, Frederik Barkhof, Mahlet A Birhanu, Esther E Bron, Robin Camarasa, Nish Chaturvedi, et al. Where is valdo? vascular lesions detection and segmentation challenge at miccai 2021. *Medical Image Analysis*, 91:103029, 2024. 5, 22
- [130] Mohammed R. S. Sunoqrot, Anindo Saha, Matin Hosseinzadeh, Mattijs Elschot, and Henkjan Huisman. Artificial intelligence for prostate mri: open datasets, available applications, and grand challenges. *European Radiology Experimental*, 6(1):35, 2022. 5, 22
- [131] David Svoboda, Michal Kozubek, and Stanislav Stejskal. Generation of digital phantoms of cell nuclei and simulation of image formation in 3d image cytometry. *Cytometry Part A: The Journal of the International Society for Advancement of Cytometry*, 75(6):494–509, 2009. 5, 22
- [132] David Svoboda, Ondřej Homola, and Stanislav Stejskal. Generation of 3d digital phantoms of colon tissue. In *Image Analysis and Recognition: 8th International Conference, ICIAR 2011, Burnaby, BC, Canada, June 22–24, 2011. Proceedings, Part II* 8, pages 31–39. Springer, 2011. 5, 22
- [133] Evropi Toulkeridou, Carlos Enrique Gutierrez, Daniel Baum, Kenji Doya, and Evan P. Economo. Automated segmentation of insect anatomy from micro-ct images using deep learning. *Natural Sciences*, 3(4):e20230010, 2023. 6, 8, 23
- [134] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H. Maier-Hein.

- Multitalent: A multi-dataset approach to medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 2023. 1
- [135] Constantin Ulrich, Tassilo Wald, Emily Tempus, Maximilian Rokuss, Paul F. Jaeger, and Klaus Maier-Hein. Radioactive: 3d radiological interactive segmentation benchmark, 2024. 6, 7, 23
- [136] Martin Vallières, Carolyn R Freeman, Sonia R Skamene, and Issam El Naqa. A radiomics model from joint fdg-pet and mri texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Physics in Medicine & Biology*, 60(14):5471, 2015. 5, 22
- [137] Jasper Willem van der Graaf, Miranda L. van Hooff, Constantinus F. M. Buckens, Matthieu Rutten, Job L. C. van Susante, Robert Jan Kroeze, Marinus de Kleuver, Bram van Ginneken, and Nikolas Lessmann. Spider - lumbar spine segmentation in mr images: a dataset and a public benchmark, 2023. 22
- [138] Bram van Ginneken. Ski10 papers, 2021. 5, 22
- [139] Kareem Wahid, Cem Dede, Mohamed Naser, and Clifton Fuller. Training dataset for hntsmrg 2024 challenge, 2024. 6, 23
- [140] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyang Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. Sam-med3d: Towards general-purpose segmentation models for volumetric medical images, 2024. 2, 4, 6
- [141] Jakob Wasserthal, Hanns-Christian Breit, Manfred Meyer, Maurice Pradella, daniel Hinck, Alexander W. Sauter, Tobias Heye, Daniel T. Boll, Joshy Cyriac, Shan Yang, Michael Bach, and Martin Segeroth. Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. *Radiol Artif Intell.*, 2023. 5, 22
- [142] Donglai Wei, Kisuk Lee, Hanyu Li, Ran Lu, J. Alexander Bae, Zequan Liu, Lifu Zhang, Márcia dos Santos, Zudi Lin, Thomas Uram, Xueying Wang, Ignacio Arganda-Carreras, Brian Matejek, Narayanan Kasthuri, Jeff Lichtman, and Hanspeter Pfister. Axonem dataset: 3d axon instance segmentation of brain cortical regions, 2021. 5, 22
- [143] Jelmer M Wolterink, Tim Leiner, Bob D De Vos, Jean-Louis Coatrieux, B Michael Kelm, Satoshi Kondo, Rodrigo A Salgado, Rahil Shahzad, Huazhong Shu, Miranda Snoeren, et al. An evaluation of automatic coronary artery calcium scoring methods with cardiac ct using the orcascore framework. *Medical physics*, 43(5):2361–2373, 2016. 5, 22
- [144] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. Scribbleprompt: Fast and flexible interactive segmentation for any biomedical image, 2024. 2, 3, 4, 5, 6
- [145] Zhaohan Xiong, Qing Xia, Zhiqiang Hu, Ning Huang, Cheng Bian, Yefeng Zheng, Sulaiman Vesal, Nishant Ravikumar, Andreas Maier, Xin Yang, et al. A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging. *Medical image analysis*, 67:101832, 2021. 5, 22
- [146] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M. Summers. Deeplesion: Automated deep mining, categorization and detection of significant radiology image findings using large-scale clinical lesion annotations, 2017. 5, 22
- [147] Jiancheng Yang, Shixuan Gu, Donglai Wei, Hanspeter Pfister, and Bingbing Ni. Ribseg dataset and strong point cloud baselines for rib segmentation from ct scans. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, 2021. 22
- [148] Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C. Paetzold, Rami Al-Maskari, Luciano Höher, Hongwei Bran Li, Ibrahim Ethem Hamamci, Anjany Sekuboyina, and et al. Benchmarking the cow with the topcow challenge: Topology-aware anatomical segmentation of the circle of willis for cta and mra, 2024. 2, 5, 22
- [149] Lin Yang, Otmar Schmid, and Fabian Isensee. Lungvis1.0: Active learning ai-powered 3d imaging ecosystem for spatial profiling of lung geometry and pulmonary nanoparticle delivery, 2023. 5, 22
- [150] Ke Zhang and Xiahai Zhuang. Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. *arXiv preprint arXiv:2203.01475*, 2022. 6
- [151] M Zhang, Y Wu, H Zhang, Y Qin, H Zheng, J Sun, GZ Yang, and Y Gu. Multi-site multi-domain airway tree modeling (atm’22). In *Proc. MICCAI Challenge*, 2022. 5, 22
- [152] Binsheng Zhao, Lawrence H Schwartz, Mark G Kris, and Gregory J Riely. Coffee-break lung ct collection with scan images reconstructed at multiple imaging parameters. In *The Cancer Imaging Archive*, 2015. 6, 23
- [153] Wei Zheng, Cheng Peng, Zeyuan Hou, Boyu Lyu, Mengfan Wang, Xuelong Mi, Shuxuan Qiao, Yinan Wan, and Guoqiang Yu. Nis3d: A completely annotated benchmark for dense 3d nuclei image segmentation. *Advances in Neural Information Processing Systems*, 36:4741–4752, 2023. 5, 22
- [154] Xiahai Zhuang, Jiahang Xu, Xinzhe Luo, Chen Chen, Cheng Ouyang, Daniel Rueckert, Victor M. Campello, Karim Lekadir, Sulaiman Vesal, Nishant RaviKumar, Yashu Liu, Gongning Luo, Jingkun Chen, Hongwei Li, Buntheng Ly, Maxime Sermesant, Holger Roth, Wentao Zhu, Jiexiang Wang, Xinghao Ding, Xinyue Wang, Sen Yang, and Lei Li. Cardiac segmentation on late gadolinium enhancement mri: A benchmark study from multi-sequence cardiac mr segmentation challenge, 2021. 6

nnInteractive: Redefining 3D Promptable Segmentation

Supplementary Material

Overview

This appendix provides more details and supporting materials to complement the main text. It includes implementation specifics of nnInteractive, training configurations, dataset descriptions, and additional experimental results.

- **Implementation Details:** Section A1 outlines preprocessing, training, and data augmentation for nnInteractive, including label instance conversion, patch sampling, and user simulation.
- **3D Bounding Box Variant of nnInteractive:** Section A2 presents details of the 3D bounding box variant developed for benchmarking, while highlighting the practicality of 2D bounding boxes.
- **Dataset:** Section A3 lists the datasets used for training and evaluation, including in-distribution (ID) datasets and challenging out-of-distribution (OOD) test datasets for assessing generalization.
- **Ambiguity:** Section A4 discusses how nnInteractive dynamically resolves segmentation ambiguities with minimal user interaction.
- **Run Time:** Section A5 presents inference speed measurements from our Napari Plugin, detailing response times across different datasets, object sizes, and interaction types.
- **Additional Results:** Section A6 provides AutoZoom’s impact on segmentation performance and qualitative comparisons demonstrating nnInteractive’s superior generalization to OOD data.

The last two images at the bottom showcase out-of-distribution data segmented using our **Napari plugin**.

A1. Implementation Details

Intensity Normalization. We perform image-level z-score normalization, meaning for each image the voxel intensities are normalized by subtracting the mean and dividing by the standard deviation.

Resampling. nnInteractive processes all images at their native resolution. No resampling to harmonize the voxels spacing is performed. This is done to ensure compatibility with out of distribution datasets where the voxel spacing may lie well outside the range common for 3D medical images.

Model training. nnInteractive is trained for a total of 5000 epochs, where each epoch is defined as 250 iterations with batch size 24. We use an input patch size of

192x192x192 voxels during training. The same patch size is also used in inference. During training, data augmentation is applied, where we extend the default nnU-Net scheme:

- **Scaling.** Scaling probability is increased from 0.2 to 0.3 per sample. With probability 0.6 we disable axis synchronization and instead sample independent scaling values for each axis. We furthermore increase the scaling range from [0.7, 1.4] to [0.5, 2]. These changes are intended to make nnInteractive less sensitive to changes in spacing (remember that we do not resample!).
- **Transpose.** With probability 0.5 we transpose random axes to simulate different image orientations
- **Intensity inversion.** With probability 0.1 we invert the intensity values of samples.

Other augmentations are unchanged, following nnU-Net’s defaults.

The remaining training settings, for example loss function, learning rate schedule etc remain identical to nnU-Net [46]. The network architecture follows ResEnc L from [47].

Training is performed on 8x Nvidia A100 40GB PCIe GPUs with a batch size of 3 per GPU. With an epoch time of about 200s, this results in total training time of 11-12 days.

Training Labels. nnInteractive exclusively performs instance segmentation, meaning that the segmentation maps encode instances of objects, rather than semantic masks. Semantic segmentations are converted to instance segmentations via connected component analysis with optionally applied morphological opening or closing depending on the dataset. nnInteractive receives no information about what semantic object a mask belongs to. Segmentations for training are stored as consecutive integers with each integer encoding one instance.

We model ambiguities by combining instances where it makes sense from a task or anatomical perspective. For example in BraTS, we use the whole tumor, tumor core, edema, enhancing and necrosis labels as targets, instead of just the edema, enhancing and necrosis.

Pseudolabels via SuperVoxels: For each training case, up to 20 SuperVoxels are generated and stored separately. Since the SuperVoxels generation may fail for very large images due to insufficient VRAM, not all training cases could be augmented with additional labels. SuperVoxels were successfully generated for 64,387/64,518 volumes. For some cases insufficient high confidence masks were generated resulting in a lower number of generated SuperVoxels. The average number of generated SuperVoxels per case is 17.46.

Patch sampling. During training patches are sampled by first randomly selecting a training case, then an object within that case and finally a pixel within that object. Sampling probabilities of training cases are determined using a mix of heuristics and manual adjustments. Three terms determine the sampling probability for each case:

- **What dataset it belongs to.** Each dataset is allocated a sampling budget equal to the square root of the number of training cases, giving datasets with more cases more representation in the training. This budget is distributed across all training cases, thus giving cases from large datasets overall a lower sampling probability than cases from smaller datasets
- **How many object each sample contains.** With the sampling probability for each case being adjusted proportional to the square root of the number of objects within them.
- **Manual dataset weights.** Datasets are up- or down weighted based on subjective perception of their usefulness for nnInteractive training. Large uniform datasets with repetitive targets (for example CAP [55] or Filo-Data3D [80]) receive lower sampling probabilities while interesting, small dataset receive higher weights (for example SegThy 1 [62]).

Once a training case was selected, a random object from that case is determined. Pseudolabels are sampled with probability 0.2. From the generated SuperVoxels one object is selected at random. For the remaining $p=0.8$ we sample real training labels. Hereby, sampling probabilities for available objects are distributed uniformly for most datasets, except for some where important classes could potentially be under-represented. For example, in Task3 from the MSD [5] we sample the liver with and without tumors with fixed probability 0.25 each, while distributing the remaining 0.5 across all present tumor instances.

Given a selected object a training patch is selected by picking a center-biased random voxel of that class and constructing the 192x192x192 patch around it.

Patches are first fed through the standard nnU-Net dataloading pipeline. At the end, an initial interaction is simulated and a user simulating agent is randomly selected.

User simulation and followup interactions. In parallel to the standard dataloading infrastructure there is a process pool for handling the user simulation. It receives prediction from the network along with the corresponding patch and previous interactions and applies the selected user agent which in turn selected an interaction to be applied for refinement (see Fig. 2).

During training, nnInteractive first checks whether a follow-up interaction is ready for processing. If it is it fetches that from the user simulation process. If not it proceeds to draw a new training sample from the standard data augmentation

pipeline. After each prediction, the current batch is given with a probability p to the user simulation for further refinement. The followup interaction probability p is initialized with 0.3 at epoch 0 and linearly increased to 0.75 during the training, unlocking more refinement steps as the model learns.

A2. 3D Bounding Box Variant of nnInteractive

The 3D bounding box version of nnInteractive was developed solely for benchmarking against existing 3D interactive segmentation models that rely on 3D bounding boxes.

In this version, 2D bounding boxes and lasso interactions were disabled, replaced entirely by 3D bounding boxes. These were simulated during training using the same methodology as 2D bounding boxes. Due to computational constraints, the model was trained for 2,000 epochs, instead of the normal 5,000 epochs, which would likely yield even better performance. Despite this, the 2K-epoch model effectively demonstrated that nnInteractive’s design principles extend well to 3D bounding boxes, outperforming all baselines by a large margin.

However, *we do not consider 3D bounding boxes a practical choice for interactive segmentation*. They are cumbersome for users to create and, in our experiments, offered no measurable advantage over 2D bounding boxes. On the contrary, they often led to false positive predictions due to the excess empty space within the bounding volume, particularly for objects oriented diagonally.

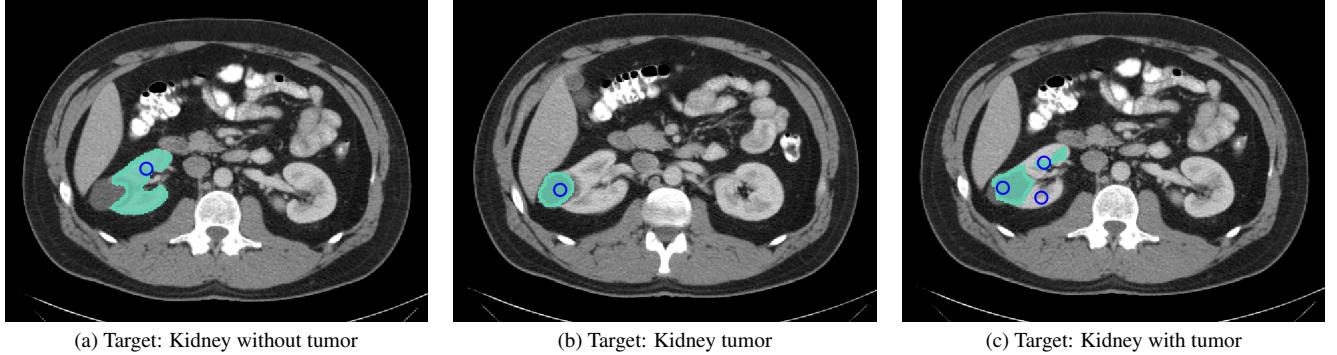
For these reasons, 3D bounding box support was intentionally excluded from the final version of nnInteractive. Beyond their inefficiency, they negatively impacted 2D bounding box and lasso performance when all interactions shared the same input channel. Due to the disparity in the number of positive input voxels, incorporating 3D bounding boxes would have required separate input channels, increasing the total input channels from 7 to 9, adding unnecessary complexity without clear benefits.

A3. Dataset

Table A1 provides an overview of all datasets used for training. Test and OOD datasets are presented in Table A2.

A4. Ambiguity

nnInteractive dynamically adapts to user input and is able to efficiently resolve ambiguities with minimal interaction Fig. A2. This is in stark contrast to competing methods that are overfitted to the training labels and lack specific ambiguity enabling training schemes (Fig. A1).

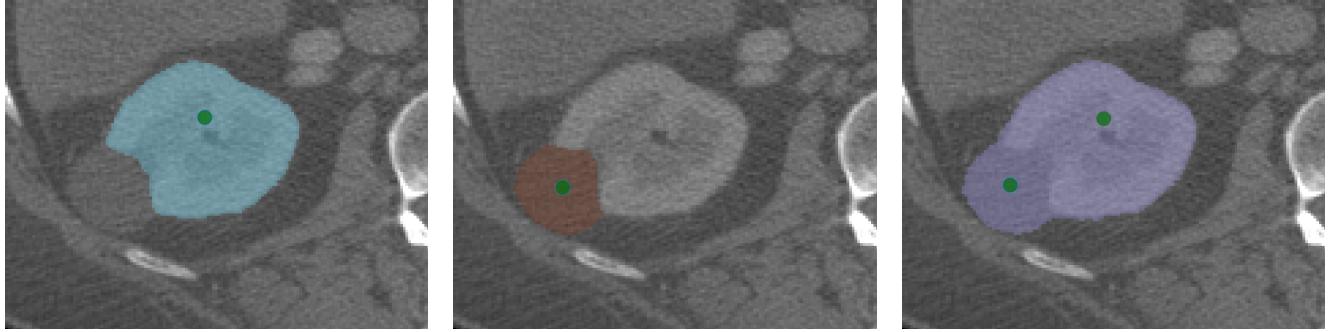


(a) Target: Kidney without tumor

(b) Target: Kidney tumor

(c) Target: Kidney with tumor

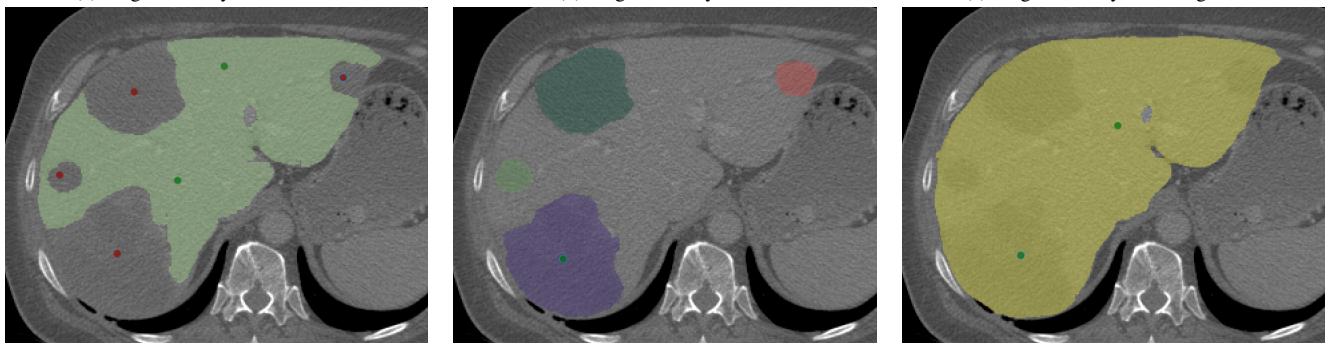
Figure A1. Vista 3D [42] cannot resolve ambiguities. When prompted to segment the kidney or the kidney tumor (a and b), Vista3D yields plausible results. However, Vista3D is unable to segment the kidney including the tumor (c), as this setting contradicts its training labels. Results generated using Nvidias online demo (<https://build.nvidia.com/nvidia/vista-3d>) and their provided *Abdomen CT* example image.



(a) Target: Kidney without tumor

(b) Target: Kidney tumor

(c) Target: Kidney including tumor



(d) Target: Liver without tumors

(e) Target: Tumors (instances)

(f) Target: Liver including tumors

Figure A2. nnInteractive successfully resolves ambiguities. Image is *liver_190* from the MSD Task 3 test set.

A5. Run Time

Table A3 shows inference times for nnInteractive as measured in our Napari Plugin. We test several images across multiple modalities, image sizes and voxel spacings. Tested objects are quite diverse ranging from known objects (organs, lymph nodes, tumors) to unknown ones (insect brain). Since nnInteractive processes objects at the original image resolution (no resampling) the size of objects in pixels is linked to the inference time. Small objects that fit within the

patch size of 192x192x192 are predicted rapidly with inference times ranging from 120-200 ms. As objects become larger, AutoZoom is required to capture them in its entirety. This raises the inference time as a function of object size. In *liver_201*, the liver requires a zoom out factor of 1.77-2.25 and 4-5 refinement boxes. Given that the time required to process a refinement box is approximately the same as processing one object without zoom, it is no surprise that the total inference type ranges rises to 1110–1290ms. As a worst case, in images such as *liver_141* inference time can

exceed 3500ms for large organs (here liver).

Row with multiple values in them indicate that several refinement steps were necessary to achieve acceptable segmentation accuracy for the respective structure. Notably, the pancreas of image *s0360* was notoriously difficult to segment. Throughout user refinement, smaller areas of the object of interest need to be changed, requiring less zoom out and consequently fewer bounding boxes for refinement. Thus, the initial interaction typically takes the longest while further interactions are much quicker.

As can be seen in the point interaction on *Dolichoderus mariae2*, the initial predictions were too localized and did not trigger AutoZoom, hinting at a deficit of point interactions in conveying informative guidance. Only after 2 additional interactions the model ‘noticed’ that the intended structure is much larger and triggered AutoZoom.

Throughout all datasets and targets we did not observe a measurable difference in inference speed across the supported interaction types.

A6. Additional results

A6.1. AutoZoom

As shown in Fig. A3, AutoZoom offers substantial benefits for large objects where high Dice scores are achieved with fewer iterations. Particularly on large objects such as livers in CT scans, AutoZoom is an essential feature.

A6.2. Qualitative Results

Additional qualitative results are shown in Fig. A4. Throughout all objects and tested prompting styles, nnInteractive consistently delivers maximum segmentation accuracy while competing methods fall behind, often producing severe artifacts. 2D methods in particular suffer from inconsistent results between slices, causing major drop in measured performance. Following nnInteractive, SegVol achieves the second best results.

nnInteractive excels even in far OOD cases such as segmenting the jaw bone of a dinosaur fossil (Fig. A5) and individual grains from sandstone (Fig. A6).



Figure A3. The effect of AutoZoom on segmentation performance. Averaged over all test and OOD datasets (a), the effect of AutoZoom is small, but noticeable. Since most target objects in these datasets are smaller than the nnInteractive patch size of 192x192x192 pixels they do not require AutoZoom, understating the impact of this contribution on large objects. When focussing on Datasets with large target objects, such as the Insect Brain of Ants (b) or the Liver in CT scans (c) AutoZoom becomes an essential feature. When AutoZoom is active, nnInteractive produces substantially improved results with fewer user interactions, achieving saturation much sooner than with AutoZoom disabled.

Table A1. Overview of the 120 datasets used for model training, covering names, image counts, modalities, targets, and access links.

| Name | Images | Modality | Target | Link |
|--------------------------------------|--------|----------------|--|---|
| Decathlon Task 2 [5, 125] | 20 | MRI | Heart | http://medicaldecathlon.com |
| Decathlon Task 3 [5, 125] | 131 | CT | Liver, L. Tumor | http://medicaldecathlon.com |
| Decathlon Task 4 [5, 125] | 208 | MRI | Hippocampus | http://medicaldecathlon.com |
| Decathlon Task 5 [5, 125] | 32 | MRI | Prostate | http://medicaldecathlon.com |
| Decathlon Task 6 [5, 125] | 63 | CT | Lung Lesion | http://medicaldecathlon.com |
| Decathlon Task 7 [5, 125] | 281 | CT | Pancreas, P. Tumor | http://medicaldecathlon.com |
| Decathlon Task 8 [5, 125] | 303 | CT | Hepatic Vessel, H. Tumor | http://medicaldecathlon.com |
| Decathlon Task 9 [5, 125] | 41 | CT | Spleen | http://medicaldecathlon.com |
| Decathlon Task 10 [5, 125] | 126 | CT | Colon Tumor | http://medicaldecathlon.com |
| ISLES2015 [91] | 28 | MRI | Stroke Lesion | http://www.isles-challenge.org/ISLES2015 |
| BTcv [67] | 30 | CT | 13 Abdominal Organs | https://www.synapse.org/Synapse:syn3193805/wiki/89480 |
| LIDC [7] | 1010 | CT | Lung Lesion | https://www.cancerimagingarchive.net/collection/lidc-idri |
| Promise12 [77] | 50 | MRI | Prostate | https://zenodo.org/records/802660 |
| ACDC [15] | 200 | MRI | RV Cavity, LV Myocardium, LV Cavity | https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html |
| ISBLesion2015 [23] | 42 | MRI | MS Lesion | https://iacl.ece.jhu.edu/index.php/MSChallenge |
| CHAOS [38] | 60 | MRI | Liver, Kidney (L and R), Spleen | https://zenodo.org/records/3431873 |
| BTcv2 [37] | 63 | CT | 9 Abdominal Organs | https://zenodo.org/records/1169361 |
| StructSeg Task1 [69] | 50 | CT | 22 OAR Head & Neck | https://structseg2019.grand-challenge.org |
| StructSeg Task2 [69] | 50 | CT | Nasopharynx Cancer | https://structseg2019.grand-challenge.org |
| StructSeg Task3 [69] | 50 | CT | 6 OAR Lung | https://structseg2019.grand-challenge.org |
| StructSeg Task4 [69] | 50 | CT | Lung Cancer | https://structseg2019.grand-challenge.org |
| SegTHOR [66] | 40 | CT | Heart, Aorta, Trachea, Esophagus | https://competitions.codalab.org/competitions/21145 |
| NIH-Pan [25, 119] | 82 | CT | Pancreas | https://wiki.cancerimagingarchive.net/display/Public/Fnacreses-CT |
| VerSe2020 [74, 81, 123] | 113 | CT | 28 Vertebrae | https://github.com/anjanv/verse |
| M&Ms [22, 93] | 300 | MRI | Left & Right Ventricle, Myocardium | https://www.uab.edu/mms |
| ProstateX [78] | 140 | MRI | Prostate Lesion | https://www.aapm.org/GrandChallenge/FROSTATEx-2 |
| RibSeg [147] | 370 | CT | Ribs | https://github.com/M3DV/RibSegTab/readme-ov-file |
| BrainMetShare [40] | 84 | MRI | Brain Metastases | https://aimi.stanford.edu/brainmetshare |
| CrossMod22 [124] | 168 | MRI | Vestibular Schwannoma, Cochlea | https://crossmod22.grand-challenge.org |
| Atlas22 [75] | 524 | MRI | Stroke Lesion | https://atlas.grand-challenge.org |
| KIT23 [43] | 489 | CT | Kidneys, K. Tumor, Cysts | https://kits-challenge.org/kits23 |
| AutoPet2 [34] | 1014 | PET,CT | Lesions | https://autopet-ii.grand-challenge.org |
| AMOS [50] | 360 | CT,MRI | 15 Abdominal Organs | https://amos22.grand-challenge.org |
| BraTS24 [8, 9, 57, 96] | 1251 | MRI | Glioblastoma | https://www.synapse.org/Synapse:syn51156910/wiki/621282 |
| AbdomenAtlas.1Mini [71, 108] | 5195 | CT | 8 Abdominal Organs | https://huggingface.co/datasets/AbdomenAtlas/_AbdomenAtlas.1Mini |
| TotalSegmentatorV2 [141] | 1180 | CT | 117 Classes of Whole Body | https://github.com/wasserth/TotalSegmentatorV2 |
| Hector2022 [4] | 524 | PET,CT | Head and Neck Tumor | https://hektor.grand-challenge.org |
| FLARE [87] | 50 | CT | 13 Abdominal Organs | https://flare22.grand-challenge.org |
| SegA [51, 104, 110] | 56 | CT | Aorta | https://multicenteraorta.grand-challenge.org/data |
| WORD [73, 84] | 120 | CT | 16 Abdominal Organs | https://github.com/HILab-git/WORD |
| AbdomenCT1K [86] | 996 | CT | Liver, Kidney, Spleen, Pancreas | https://github.com/JunMail/AbdomenCT-1K |
| DAP-ATLAS [48] | 533 | CT | Lung, Brain, Bones, Liver, Kidney, Bladder | https://github.com/alexanderjaus/AtlasDataset |
| CTORG [114] | 140 | CT | Vessel Components of CoW | https://www.cancerimagingarchive.net/collection/ct-org |
| TopCow [148] | 200 | CT,MRI | Lungs, Airways | https://topcow23.grand-challenge.org |
| AortaSeg24 [45] | 50 | CT | Axon Instances | https://aortaseg24.grand-challenge.org |
| Duke Liver [90] | 310 | MRI | Axon Instances | https://zenodo.org/records/7774566 |
| Aero Path [128] | 27 | CT | Mitochondria Instances | https://github.com/raidenics/AeroPath |
| AxonEM [142] | 18 | El. Microscopy | Neuronal Nuclei | https://axonem.grand-challenge.org |
| MitoEM [142] | 4 | El. Microscopy | Airway | https://mitoem.grand-challenge.org |
| NucMM [76] | 62 | El. Microscopy | Cell Nuclei | https://nucmm.grand-challenge.org |
| LungVis1.0 [149] | 22 | Fl. Microscopy | Colon Tissue | https://zenodo.org/records/7413818 |
| BBBC024 HL60 Cell line [131] | 240 | Fl. Microscopy | Blastocyst Cells | https://bbbc.broadinstitute.org/BBBC024 |
| BBBC027 Colon Tissue [132] | 60 | Fl. Microscopy | Trophoblast | https://bbbc.broadinstitute.org/BBBC027 |
| BBBC032 Mouse Embryo/Blastocyst [80] | 1 | Fl. Microscopy | Stem Cells | https://bbbc.broadinstitute.org/BBBC032 |
| BBBC033 C16 Trophoblast [80] | 1 | Microscopy | Lung Cancer Cells | https://bbbc.broadinstitute.org/BBBC033 |
| BBBC034 Pluripotent Stem Cells [80] | 1 | Microscopy | Mouse Embryonic Cells | https://bbbc.broadinstitute.org/BBBC034 |
| BBBC046 FiloData3D [80] | 5400 | Fl. Microscopy | Endocardium, Epicardium, Atrium | https://bbbc.broadinstitute.org/BBBC046 |
| BBBC050 Mouse Embryo/Nuclei [80] | 165 | Fl. Microscopy | LV Lumen | https://bbbc.broadinstitute.org/BBBC050 |
| CAMUS [68] | 1000 | US | Mitochondria | https://www.creatis.insa-lyon.fr/Challenge/camus/index.html |
| CETUS [14] | 90 | US | Brain Regions | https://www.creatis.insa-lyon.fr/Challenge/CETUS/databases.html |
| EPEI_Mito [82] | 1 | El. Microscopy | Mitochondria | https://www.epfl.ch/labs/cvlab/data/data-em |
| FETA [102] | 120 | MRI | Mitochondria, Synapses | https://fetachallenge.github.io/pages/data_description |
| Drosophila [36] | 1 | El. Microscopy | Lower-Limb Leg | https://github.com/unidesigner/grounder-drosophila-vnc |
| Leg3DUS [31] | 44 | US | Tumor | https://www.cs.cmu.edu/camp/publications/leg-3d-us-dataset |
| LGGMRISeg [21] | 110 | MRI | Neonatal Brain Atlas | https://www.kaggle.com/datasets/mateuszbdz/lgg-mri-segmentation/data |
| M-CRIB [3] | 10 | MRI | Mineral Samples | https://osf.io/4vthr |
| ParticleSeg3D [39] | 54 | MicroCT | Cerebral Tumor | https://syncandshare.desy.de/index.php/s/wj1D049KangiPj5 |
| RESECT [12] | 69 | US | Left Ventricle | https://www.cardiacatlases.org/lv-segmentation-challenge |
| CAP [55] | 1637 | MRI | Left Atrium | https://zenodo.org/records/11456029 |
| AtriaSeg2018 [145] | 100 | MRI | Cell Nuclei | https://www.cs.cmu.edu/camp/publications/seghy-dataset |
| NIS3D [153] | 6 | Fl. Microscopy | Cell, Border | https://www.cs.cit.tum.de/camp/publications/seghy-dataset |
| SegThy 1 [62] | 14 | MRI | Vertebra | https://celltrackingchallenge.net/3d-datasets |
| SegThy 2 [62] | 32 | US | White Matter Hyperintensities | https://www.cancerimagingarchive.net/collection/spine-mts-ct-seg |
| Fluo C3DH A549 [94] | 90 | Fl. Microscopy | Prostate | https://www.cancerimagingarchive.net/analysis-result/lsebi-mr-prostate-2013 |
| Fluo N3DH [94] | 230 | Fl. Microscopy | Brain Regions | https://sites.wustl.edu/oaislbrains/home/oasis-1 |
| Spine-Mets [105] | 55 | CT | Mediastinal Lymph Nodes | https://github.com/dbouget/ct_mediastinal_structures_segmentation |
| WMHSegChallenge [64] | 60 | MRI | Mediastinal Structures | https://github.com/dbouget/ct_mediastinal_structures_segmentation |
| NCI-ISBI [16] | 59 | MRI | Lymph Nodes | https://www.cancerimagingarchive.net/collection/ct-lymph-nodes |
| OASIS [92] | 436 | MRI | Breast Lesions | https://www.synapse.org/Synapse:syn6086042/wiki/628716 |
| MediaLymph [19] | 15 | CT | Airway Tree | https://atm22.grand-challenge.org |
| MediaStruct [18] | 15 | CT | Organs | https://doi.org/10.7937/TCIA.X0HO-1706 |
| CT Lymp Nodes [118] | 175 | CT | Cervix, Tumor | https://atlas-challenge.u-bourgogne.fr |
| MAMA MIA [33] | 1506 | MRI | Pancreas, Kidney, Liver | https://www.curvas.grand-challenge.org/curvas-dataset |
| ATM2022 [151] | 300 | CT | Heart Structures | https://emidec.com |
| Pediatric CT SEG [52] | 353 | CT | Kidney, Vessel, Tumor | https://segchd.csail.mit.edu |
| Atlas Bourgogne [109] | 60 | MRI | Heart, Vessel | https://kipa22.grand-challenge.org |
| CC Tumor Heterogeneity [95] | 63 | MRI | Brain Structures | https://mbrains18.isi.uu.nl/index.html |
| CURVAS [113] | 60 | CT | Califications | https://parse2022.grand-challenge.org/Parse2022 |
| Emidec [65] | 100 | MRI | Pulmonary Artery | https://www.imagenglab.com/newsite/pddca |
| HVSMDR-2.0 [101] | 60 | MRI | Bladder, Prostate, Rectum | https://www.cancerimagingarchive.net/collection/prostate-anatomical-edge-cases |
| Kipa22 [41] | 70 | CT | Cartilage, Bone | https://sk10.grand-challenge.org |
| MrBrains18 [63] | 30 | MRI | Edema, Tumor | https://zenodo.org/records/10159290 |
| OrCaScore [143] | 32 | CT | Lumbar Spine | https://zenodo.org/records/11367005 |
| Parse22 [83] | 100 | CT | Liver, Tumor | https://instancex.grand-challenge.org |
| PDPCA [111] | 47 | CT | Edema, Tumor | https://ditte.ing.unimore.it/toothfairy2 |
| ProstateEdgeCases [56] | 131 | CT | Edema, Tumor | https://www.cancerimagingarchive.net/collection/ct-lymph-nodes |
| SK10 [138] | 100 | MRI | Edema, Tumor | https://zenodo.org/records/6481141 |
| Soft Tissue Sarcoma [136] | 102 | MRI | Edema, Tumor | https://zenodo.org/records/10159290 |
| Spider [137] | 447 | MRI | Edema, Tumor | https://zenodo.org/records/11367005 |
| VALDO Task 2 [129] | 72 | MRI | Edema, Tumor | https://valdo.grand-challenge.org/Task2 |
| ToothFairy 2 [17] | 480 | CT | Edema, Tumor | https://www.cancerimagingarchive.net/collection/opennn-gbm |
| UPENN-GBM [10] | 147 | MRI | Edema, Tumor | https://www.cancerimagingarchive.net/collection/remin |
| ReMiND [54] | 213 | MRI | Edema, Tumor | https://zenodo.org/records/6481141 |
| Prostate158 [130] | 188 | MRI | Edema, Tumor | https://zenodo.org/records/10159290 |
| TotalSegmentator MRI [26] | 298 | MRI | Edema, Tumor | https://instancex.grand-challenge.org |
| Instance2022 [72] | 100 | CT | Intracranial Hemorrhage | https://cebs-ext.niehs.nih.gov/cahs/report/lapd/web-download-links |
| LAPD Mouse [13] | 34 | Fl. Microscopy | Airway | https://zenodo.org/records/3757476 |
| Deep Lesion [146] | 1093 | CT | Multiple Types of Lesions | https://nihcc.app.box.com/v/DeepLesion |
| COVID-19 CT Lung [53] | 10 | CT | COVID-19 | https://zenodo.org/records/11367005 |
| LNDb [103] | 229 | CT | Lymph Nodes | https://lndb.grand-challenge.org |
| NIH Lymph [117] | 176 | CT | Lymph Nodes | https://www.cancerimagingarchive.net/collection/ct-lymph-nodes |
| NSCLC Pleural Effusion [61] | 78 | CT | Pleural Effusion | https://www.cancerimagingarchive.net/collection/ct-lymph-nodes |
| NSCLC Radiomics [2] | 503 | CT | Lung Lesions | https://www.cancerimagingarchive.net/collection/nsclc-radiomics |
| COVID-19-20 [120] | 199 | CT | COVID-19 | https://covid-segmentation.grand-challenge.org/COVID-19-20 |

Table A2. **Overview of held-out test datasets** spanning diverse anatomical structures, pathologies, and imaging modalities. These datasets introduce significant domain shifts in resolution, contrast, target structures, and anatomical scale, serving as a rigorous benchmark for model robustness. We use the filtered datasets from the RadioActive benchmark [135] along with four additional out-of-distribution (OOD) datasets, visually inspected before benchmarking. None were part of the training data, except for SegVol, which included HanSeg.

| | Dataset | Modality | Targets | Images |
|--|---------------------|-----------------|---|---------------|
| RadioActive Benchmark Datasets [135] | MS Lesion [99] | MRI (T2 Flair) | MS Lesions | 60 |
| | HanSeg [106] | MRI (T1) | 30 Organs at Risk | 42 |
| | HNTSRMFG [139] | MRI (T2) | Oropharyngeal Cancer and Metastatic Lymph Nodes | 135 |
| | RiderLung [152] | CT | Lung Lesions | 58 |
| | LNQ [29] | CT | Mediastinal Lymph Nodes | 513 |
| | LiverMets [126] | CT | Liver Metastases | 171 |
| | Adrenal ACC [98] | CT | Adrenal Tumors | 53 |
| | HCC Tace [97] | CT | Liver and Liver Tumors | 65 |
| | Pengwin [79] | CT | Bone Fragments | 100 |
| | SegRap [85] | CT | 45 Organs at Risk | 30 |
| Additional OOD Datasets | MouseTumor [49] | MicroCT | Subcutaneous Tumors in Mice | 452 |
| | InsectAnatomy [133] | MicroCT | Insect Brain | 84 |
| | ACRIN H&N [59] | PET | Head and Neck Tumors | 67 |
| | Stanford Knee [28] | MRI | Patellar, Femoral, Tibial Cartilages and Meniscus | 155 |

Table A3. Inference run times. Measured using our Napari Plugin on a system with an Nvidia RTX 4090, AMD Ryzen 5800X3D and 32GB RAM. Multiple numbers per box indicate several refinement steps by the user. Inference time is measured as wall time from triggering the prediction until the prediction is completed. This includes copying data between devices, and all steps of the inference pipeline, including the computation of zoom steps, determining what boxes to use for refinement, etc.

| Image | Target | Interaction | Zoom out | Refinement Boxes | Time AutoZoom (ms) | Time Refinement (ms) | Time Total (ms) |
|---|------------------|-------------|---------------|------------------|--------------------------|----------------------|-------------------------|
| Image: Dolichoderus_mariae2 Dataset: InsectAnatomy Modality: MicroCT Size: 260x260x99 | Insect Brain | lasso | 1.38 | 5 | 240 | 610 | 850 |
| | | scribble | 1.5 | 5 | 360 | 700 | 1050 |
| | | 2D bbox | 1.39 | 6 | 160 | 890 | 1050 |
| | | point | 1; 1; 2.25; 1 | 0; 0; 4; 0 | 130; 120; 450; 150 | 0; 0; 540; 0 | 130; 120; 990; 150 |
| Image: lq_0006 Dataset: LNQ Modality: CT Size, spacing: 512x512x118, 0.9x0.9xmm | Lymph Node | lasso | 1 | 0 | 160 | 0 | 160 |
| | | scribble | 1 | 0 | 180 | 0 | 180 |
| | | 2D bbox | 1 | 0 | 150 | 0 | 150 |
| | | point | 1 | 0 | 160 | 0 | 160 |
| Image: Mets_019 Dataset: BrainMetShare Modality: T1 spin-echo post Size, spacing: 256x256x133, 0.94x0.94x1mm | Brain Metastasis | lasso | 1 | 0 | 160 | 0 | 160 |
| | | scribble | 1 | 0 | 140 | 0 | 140 |
| | | 2D bbox | 1 | 0 | 130 | 0 | 130 |
| | | point | 1 | 0 | 150 | 0 | 150 |
| Image: s0360 Dataset: Totalsegmentator MRI v2 Modality: MRI Size, spacing: 512x512x96, 0.72x0.72x1.7mm | Liver | lasso | 1.59 | 4 | 190 | 580 | 770 |
| | | scribble | 1.73 | 5 | 340 | 720 | 1060 |
| | | 2D bbox | 1.54 | 5 | 190 | 720 | 910 |
| | | point | 1.5 | 5 | 380 | 660 | 1040 |
| | Spleen | lasso | 1.39 | 3 | 160 | 450 | 610 |
| | | scribble | 1.52 | 3 | 290 | 440 | 740 |
| | | 2D bbox | 1.36 | 3 | 170 | 400 | 570 |
| | | point | 1.5 | 3 | 380 | 460 | 850 |
| | Aorta | lasso | 1 | 0 | 170 | 0 | 170 |
| | | scribble | 1 | 0 | 150 | 0 | 150 |
| | | 2D bbox | 1 | 0 | 150 | 0 | 150 |
| | | point | 1 | 0 | 160 | 0 | 160 |
| | Pancreas | lasso | 1.5; 1.5; 1 | 3; 1; 0 | 320; 310; 200 | 450; 130; 0 | 770; 440; 200 |
| | | scribble | 1.5; 1.5; 1 | 3; 2; 0 | 290; 270; 190 | 360; 260; 0 | 650; 540; 190 |
| | | 2D bbox | 1.5; 1; 1 | 3; 0; 0; 0 | 290; 130; 170; 130 | 360; 0; 0; 0 | 650; 130; 170; 130 |
| | | point | 2.25; 1; 1; 1 | 1; 0; 0; 0; 0 | 620; 150; 120; 150; 130; | 210; 0; 0; 0 | 830; 150; 120; 150; 130 |
| | Kidney L | lasso | 1 | 0 | 170 | 0 | 170 |
| | | scribble | 1 | 0 | 150 | 0 | 150 |
| | | 2D bbox | 1 | 0 | 160 | 0 | 160 |
| | | point | 1 | 0 | 160 | 0 | 160 |
| Image: liver_201 Dataset: Task 3 MSD Modality: CT Size, spacing: 512x512x186, 0.77x0.77x2.5mm | Liver | lasso | 1.91 | 5 | 400 | 760 | 1160 |
| | | scribble | 2.21 | 5 | 560 | 660 | 1210 |
| | | 2D bbox | 1.77 | 5 | 330 | 780 | 1110 |
| | | point | 2.25 | 4 | 740 | 540 | 1290 |
| | Spleen | lasso | 1 | 0 | 130 | 0 | 130 |
| | | scribble | 1 | 0 | 190 | 0 | 190 |
| | | 2D bbox | 1 | 0 | 160 | 0 | 160 |
| | | point | 1 | 0 | 180 | 0 | 180 |
| | Vertebra | lasso | 1 | 0 | 170 | 0 | 170 |
| | | scribble | 1 | 0 | 160 | 0 | 160 |
| | | 2D bbox | 1 | 0 | 170 | 0 | 170 |
| | | point | 1 | 0 | 160 | 0 | 160 |
| | Pancreas | lasso | 1 | 0 | 160 | 0 | 160 |
| | | scribble | 1 | 0 | 220 | 0 | 220 |
| | | 2D bbox | 1.04 | 1 | 210 | 190 | 400 |
| | | point | 1.5 | 1 | 390 | 170 | 560 |
| | | lasso | 1 | 0 | 160 | 0 | 160 |
| | Kidney L | scribble | 1 | 0 | 180 | 0 | 180 |
| | | 2D bbox | 1 | 0 | 170 | 0 | 170 |
| | | point | 1 | 0 | 170 | 0 | 170 |
| Image: liver_141 Dataset: MSD Task 3 Modality: CT Size, spacing: 512x512x971 0.7x0.7x0.5mm | Liver | lasso | 2.91 | 14 | 1520 | 2180 | 3700 |
| | | scribble | 3.54 | 14 | 1650 | 1890 | 3540 |
| | | 2D bbox | 2.69 | 14 | 1300 | 2300 | 3600 |
| | | point | 3.36; 1.5 | 7; 6 | 1980; 500 | 1160; 980 | 3150; 1490 |
| | Spleen | lasso | 1.63 | 4 | 500 | 640 | 1130 |
| | | scribble | 1.5 | 4 | 530 | 640 | 1180 |
| | | 2D bbox | 1.67 | 3 | 600 | 570 | 1170 |
| | | point | 1.5 | 4 | 490 | 650 | 1140 |
| | Vertebra | lasso | 1 | 0 | 180 | 0 | 180 |
| | | scribble | 1 | 0 | 150 | 0 | 150 |
| | | 2D bbox | 1 | 0 | 260 | 0 | 260 |
| | | point | 1 | 0 | 200 | 0 | 200 |
| | Pancreas | lasso | 1.5 | 2 | 570 | 340 | 910 |
| | | scribble | 1.5 | 2 | 590 | 370 | 860 |
| | | 2D bbox | 1 | 1 | 190 | 0 | 190 |
| | | point | 1 | 1 | 190 | 0 | 190 |
| | Kidney L | lasso | 1.5 | 3 | 490 | 560 | 1020 |
| | | scribble | 1.5 | 3 | 510 | 560 | 1070 |
| | | 2D bbox | 1.5 | 3 | 510 | 650 | 1130 |
| | | point | 1.5 | 3 | 500 | 510 | 1010 |

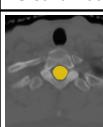
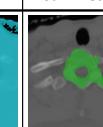
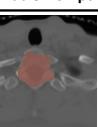
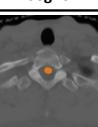
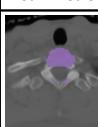
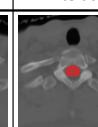
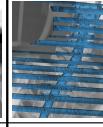
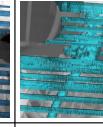
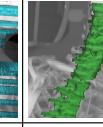
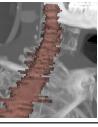
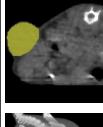
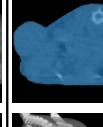
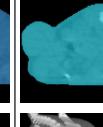
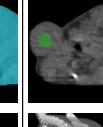
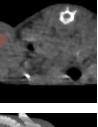
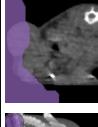
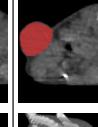
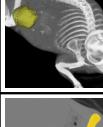
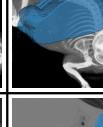
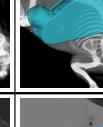
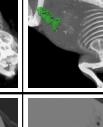
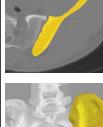
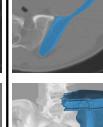
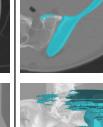
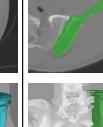
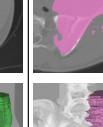
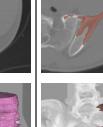
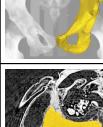
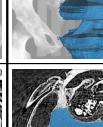
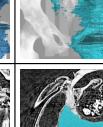
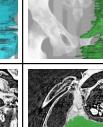
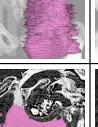
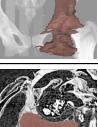
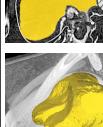
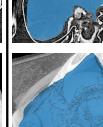
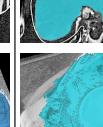
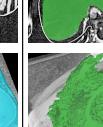
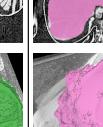
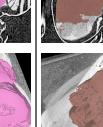
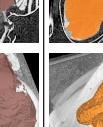
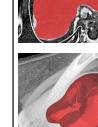
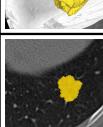
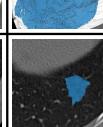
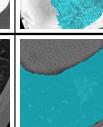
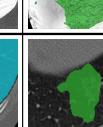
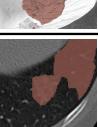
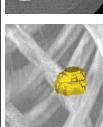
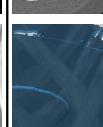
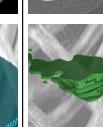
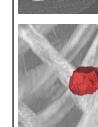
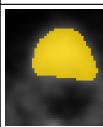
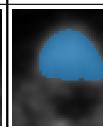
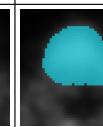
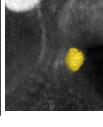
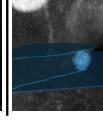
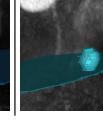
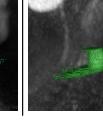
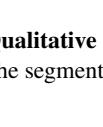
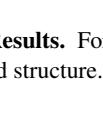
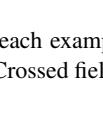
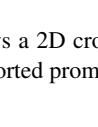
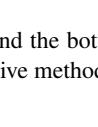
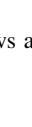
| | Ground Truth | SAM | SAM2 | SamMed 2D | MedSam | ScribblePrompt | SegVol | SamMed 3D | nnInteractive |
|-------------------|---|---|---|---|---|---|---|---|---|
| Single Point |  |  |  |  | X |  |  |  |  |
| |  |  |  |  |  | X |  |  |  |
| Single Point |  |  |  |  | X |  |  |  |  |
| |  |  |  |  |  | X |  |  |  |
| Bounding Box (3D) |  |  |  |  |  |  |  | X |  |
| |  |  |  |  |  | X |  |  |  |
| Bounding Box (3D) |  |  |  |  |  |  |  | X |  |
| |  |  |  |  |  | X |  |  |  |
| Scribble |  |  |  |  | X |  |  | X |  |
| |  |  |  |  | X |  | X |  |  |
| Scribble |  |  |  |  | X |  |  | X |  |
| |  |  |  |  | X |  | X |  |  |

Figure A4. **Qualitative Results.** For each example the top row shows a 2D cross-section of the target and the bottom row shows a 3D rendering of the segmented structure. Crossed fields X denote unsupported prompting style for the respective method

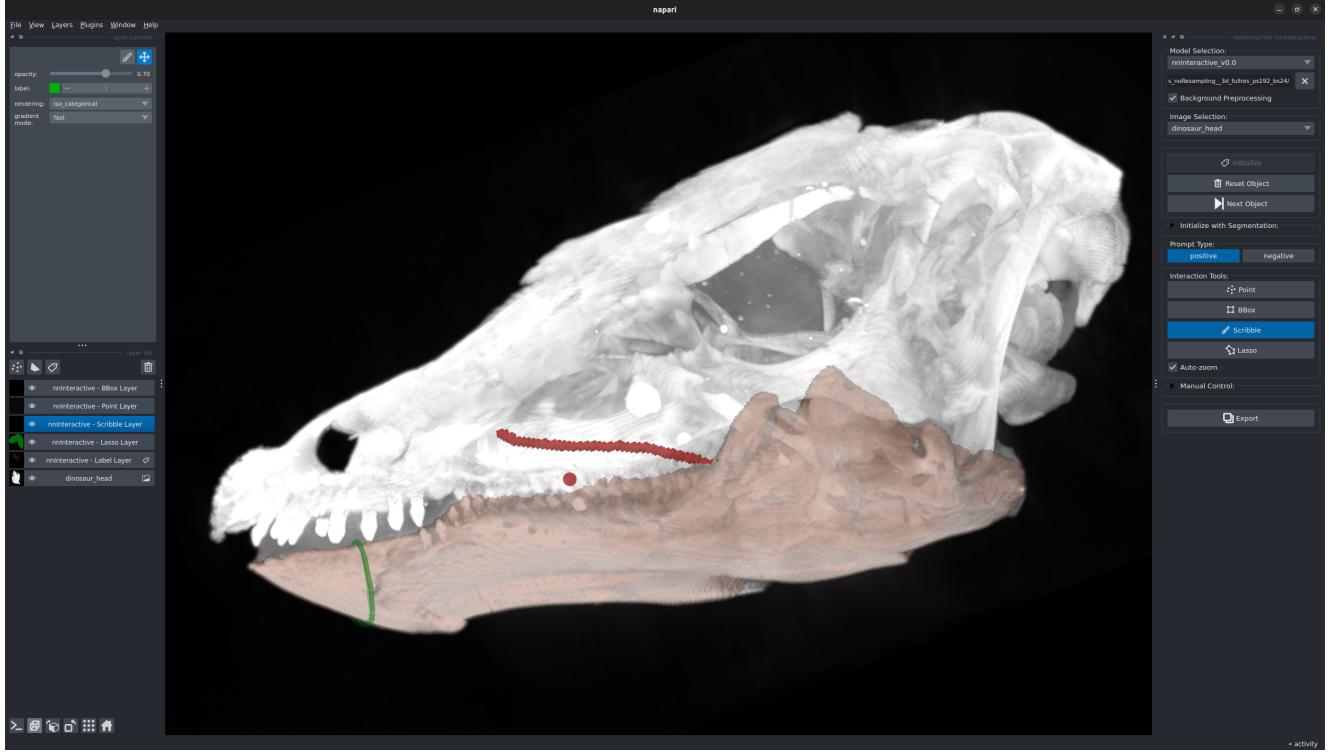


Figure A5. nnInteractive on out-of-distribution data. Segmentation of dinosaur (*Thescelosaurus neglectus* [20]) jawbones using the proposed Napari plugin with a few intuitive prompts (lasso, scribbles, points). Green areas indicate positive prompts while dark red highlights negative prompting. This demonstrates nnInteractive’s ability to generalize to unseen, non-medical data.

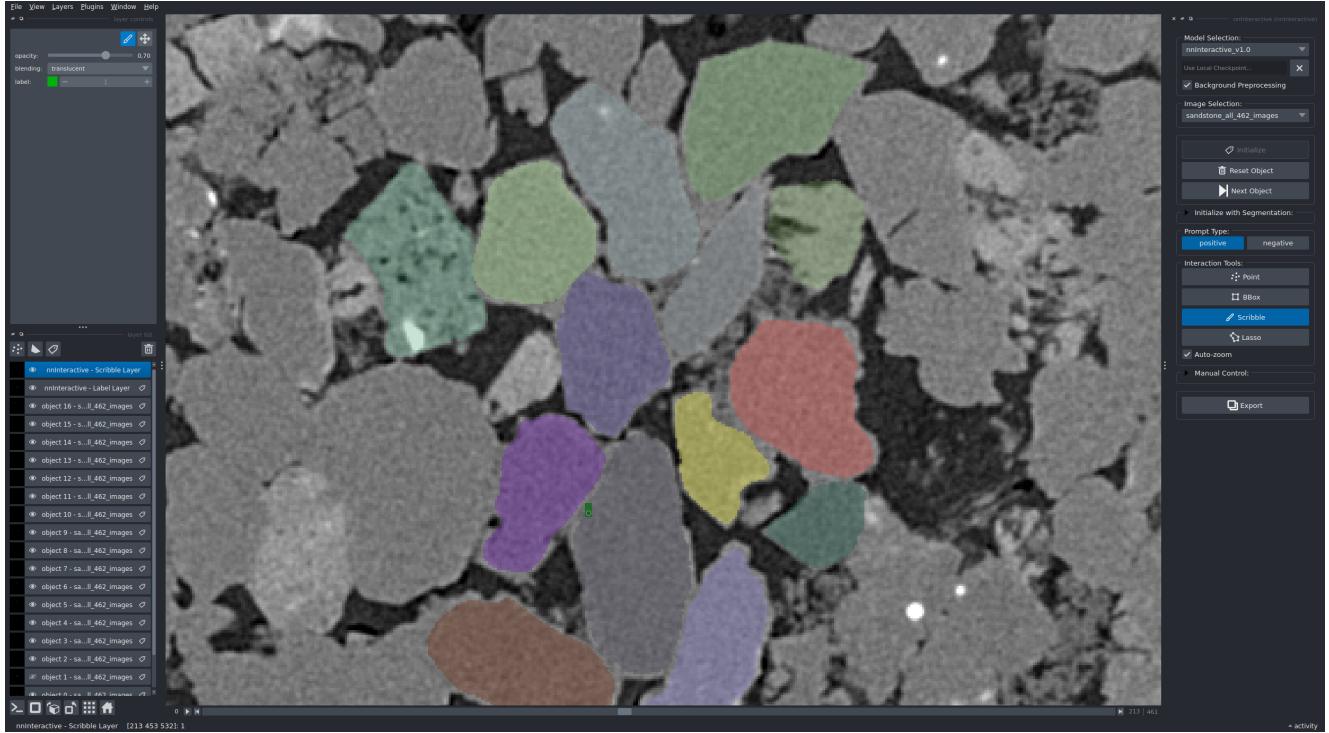


Figure A6. nnInteractive on out-of-distribution data. MicroCT of sandstone (data source: [here](#)). nnInteractive successfully separates individual grains despite the borders being barely visible to the human eye.