Karlsruher Institut für Technologie

INTERACTIVE SEGMENTATION AND ANNOTATION OF MEDICAL IMAGES

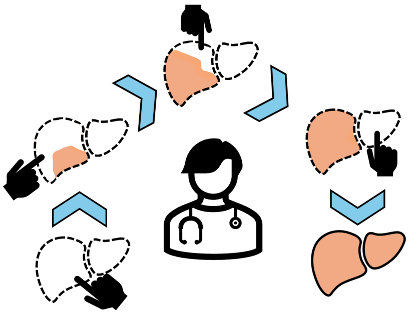
Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

von der KIT-Fakultät für Informatik des
Karlsruher Institut für Technologie (KIT)
genehmigte

Dissertation von
ZDRAVKO MARINOV

geboren in Sofia, Bulgarien



Tag der mündlichen Prüfung: 05. Februar 2026

Hauptreferent: Prof. Dr.-Ing. Rainer Stiefelhagen
Korreferent: Prof. Dr. med. Dr. rer. nat. Jens Kleesiek
Korreferent: Prof. Maria Zuluaga

Zdravko Marinov: *Interactive Segmentation and Annotation of Medical Images*

„Сговорна дружина, планина повдига.“
“A well-united group can lift a mountain.”

— Bulgarian proverb

Dedicated to my little sunshine – Annika.

ABSTRACT

Medical image segmentation has advanced rapidly, with models achieving human-level performance on many benchmarks. Yet, expert oversight remains essential, as imperfect predictions can directly influence diagnosis and treatment. Interactive segmentation addresses this challenge by integrating human feedback into models, enabling users to iteratively refine predictions through clicks, scribbles, or bounding boxes. This allows experts to steer models toward accurate results and efficiently annotate complex data, such as 3D volumes, with a few interactions, rather than laborious voxel-wise annotations. Despite its promise, the field remains fragmented - lacking clear definitions, standardized evaluation, realistic interaction simulation, and unified benchmarks. This thesis tackles these core challenges, providing both practical solutions and theoretical foundations that have been largely absent in prior research.

In the first part of this thesis, we examine three core components of interactive models and establish best practices applicable to any model regarding: (1) the representation of human interactions; (2) the realistic simulation of medical annotators' interactions; and (3) the efficiency of interactive models. We provide formal definitions for interaction representation ("guidance signal") and interaction simulation ("robot user"), and introduce a theoretical framework to systematically describe interactive models using our terminology.

Building on these theoretical foundations, in the second part of this thesis, we conduct a systematic review of 121 existing interactive segmentation methods and construct a taxonomy that classifies methods into distinct categories. Using our taxonomy, we provide practical recommendations on model selection for specific applications and outline best practices for each taxonomy category. Our review also reveals major gaps in the field, including the lack of standardized evaluation, datasets, baselines, and interaction simulation protocols, which have caused many prior models to be evaluated in isolation without comparison to existing work.

To address the lack of common standards, we establish a global interactive segmentation community by co-organizing four international competitions across multiple centers, attracting over 200 submissions. These efforts allow a community-wide discussion of key research gaps and provide a foundation for fair and standardized comparison of interactive approaches. We also compile a large-scale multimodal dataset of 166 open datasets across nine imaging modalities, with standardized protocols for interaction simulation and evaluation, enabling transparent and reproducible benchmarking in any medical imaging domain.

This thesis redefines the foundations of interactive medical image segmentation through three pillars: theoretical contributions, practical recommendations, and community-driven benchmarks. These foundations provide the structure needed for progress, showing that advancement depends on collaboration, establishing best practices, and identifying research gaps to move the field forward toward clinically reliable, annotation-efficient interactive models.

ZUSAMMENFASSUNG

Die medizinische Bildsegmentierung hat sich in den letzten Jahren rasant weiterentwickelt, und viele Modelle erreichen inzwischen eine dem Menschen vergleichbare Leistungsfähigkeit auf zahlreichen Benchmarks. Dennoch bleibt die Expertise von Fachpersonen unverzichtbar, da unvollständige oder fehlerhafte Vorhersagen unmittelbaren Einfluss auf Diagnose und Therapie haben können. Interaktive Segmentierung begegnet dieser Herausforderung, indem sie menschliches Feedback in Modelle integriert und es Nutzerinnen ermöglicht, Vorhersagen iterativ über menschliche Interaktionen wie Klicks zu verfeinern. So können Expertinnen Modelle gezielt zu präzisen Ergebnissen leiten und komplexe Daten – etwa 3D-Volumina – mit wenigen Interaktionen statt durch aufwendige voxelweise Annotationen effizient annotieren. Trotz dieses Potenzials ist das Forschungsfeld fragmentiert: Es fehlen klare Definitionen, standardisierte Evaluationsverfahren, realistische Interaktionssimulationen und einheitliche Benchmarks. Diese Dissertation widmet sich diesen zentralen Herausforderungen und liefert sowohl praktische Lösungen als auch theoretische Grundlagen, die bisher weitgehend gefehlt haben.

Im ersten Teil der Dissertation untersuchen wir drei zentrale Komponenten interaktiver Modelle und formulieren Best Practices, die für jedes Modell anwendbar sind: (1) die Repräsentation menschlicher Interaktionen; (2) die realistische Simulation der Interaktionen medizinischer Annotator*innen; (3) sowie die Effizienz interaktiver Modelle. Wir geben formale Definitionen für die Interaktionsrepräsentation („guidance signal“) und die Interaktionssimulation („robot user“) und entwickeln einen theoretischen Rahmen, um interaktive Modelle systematisch mit unserer Terminologie zu beschreiben.

Aufbauend auf diesen theoretischen Grundlagen führen wir im zweiten Teil der Dissertation eine systematische Analyse von 121 existierenden Methoden der interaktiven Segmentierung durch und entwickeln eine Taxonomie, die die Methoden in klar abgegrenzte Kategorien einordnet. Mithilfe dieser Taxonomie geben wir praktische Empfehlungen zur Modellauswahl für spezifische Anwendungen und formulieren Best Practices für jede Kategorie. Unsere Analyse zeigt außerdem wesentliche Lücken im Forschungsfeld auf, darunter das Fehlen standardisierter Evaluationen, Datensätze, Baselines und Protokolle zur Interaktionssimulation. Diese Defizite haben dazu geführt, dass viele frühere Modelle isoliert evaluiert wurden – ohne Vergleich zu bestehender Arbeit.

Um das Fehlen gemeinsamer Standards zu adressieren, etablieren wir eine globale Community für interaktive Segmentierung, indem wir vier internationale Wettbewerbe an mehreren Standorten mitorganisieren, die über 200 Einreichungen anziehen. Diese Initiativen ermöglichen eine gemeinschaftsweite Diskussion zentraler Forschungslücken und schaffen die Grundlage für faire und standardisierte Vergleiche interaktiver Ansätze. Zudem erstellen wir ein groß angelegtes, multimodales Datenset aus 166 offenen Datensätzen über neun Bildgebungsmodalitäten, ergänzt durch standardisierte Protokolle für Interaktionssimulation und Evaluation, sodass transparente und reproduzierbare Benchmarks in jedem Bereich der medizinischen Bildgebung möglich werden.

Diese Dissertation definiert die Grundlagen der interaktiven medizinischen Bildsegmentierung neu – basierend auf drei Säulen: theoretischen Beiträgen, praktischen Empfehlungen und Community-getriebenen Benchmarks. Diese Grundlagen schaffen die notwendige Struktur für Fortschritt und zeigen, dass Weiterentwicklung von Zusammenarbeit, der Etablierung von Best Practices und der Identifikation zentraler Forschungslücken abhängt, um das Feld in Richtung klinisch verlässlicher und annotierungseffizienter interaktiver Modelle voranzubringen.

ACKNOWLEDGMENTS

The work I carried out over the past three and a half years at CV:HCI, culminating in this dissertation, would not have been possible without the steady support, motivation, and inspiration of many people - my advisors, mentors, colleagues, friends, and family.

First and foremost, I would like to thank Prof. Rainer Stiefelhagen for his outstanding guidance throughout my time at KIT. I am truly grateful not only for his contributions to my research, but also for the trust and independence he gave me to explore new ideas, paired with his insightful feedback whenever I needed direction. His balance of freedom and mentorship created the perfect environment for developing my own research path.

My sincere gratitude also goes to my second advisor, Prof. Jens Kleesiek. His belief in my work and his encouragement to aim high, especially when submitting to major conferences, meant a great deal to me. I also deeply appreciate his willingness to help me navigate the medical domain and his openness to unconventional, sometimes bold ideas. His feedback consistently helped me understand what truly matters in clinical practice.

I am equally honoured to have Prof. Maria Zuluaga as my third reviewer. Her pioneering contributions to medical interactive segmentation have shaped the field in remarkable ways and have inspired me since the earliest stages of my research. Having her involved in this dissertation is a privilege, and I hope that our research paths will cross again in the future.

One of the greatest joys of my PhD journey has been the incredible community at CV:HCI. I owe thanks to Alexander, Constantin, Corinna, Simon, Jiaming, Yufan, Kunyu, Jiale, Omar, David, Alina, Junwei, Ruiping, Di, and Saquib for creating an enjoyable, supportive, and intellectually stimulating atmosphere. Working with them made everyday life at the lab a pleasure.

I am also grateful for HIDSS4Health for their continuous support, in particular to Nicole, who has always been dedicated and supportive. My heartfelt appreciation goes to the challenge co-organizers who put their trust in me and supported the interactive segmentation initiative: Thomas, Federico, Jun, Marawan, and all others who contributed along the way.

To my family and friends: thank you for being my anchor outside the academic world. Your encouragement, humour, and patience kept me grounded throughout this journey. I am especially thankful to my parents, Binka and Todor, and to Kristina and Hristo, along with my nephew Kaloyan, for their constant love and support. To my closest friends: Rozi, Boyko, Kris, Martin, Alex, Vasko, Tosho, Gosho, Tseko, Nasko, Stuci, Sashks, Eli, Yavor, Sonya, and Juji, thank you for making life brighter and for always being there, even when I disappeared into research for months at a time.

And lastly, from the bottom of my heart, I want to thank Stanka and Annika. You bring light and warmth into my everyday life, and your love makes all challenges easier to face. You are my two rays of sunshine, and I am endlessly grateful for you.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the Curve
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CT	Computed Tomography
DICOM	Digital Imaging and Communications in Medicine
DICOM-SEG	DICOM Segmentation
DMF	Distance Maps Fusion
DSC	Dice Similarity Coefficient
EDT	Euclidean Distance Transform
FDG	Fluorodeoxyglucose
FLAIR	Fluid-Attenuated Inversion Recovery
FTP	File Transfer Protocol
GDT	Geodesic Distance Transform
GPU	Graphics Processing Unit
H5	Hierarchical Data Format 5

HTTP	Hypertext Transfer Protocol
IAC	Inferior Alveolar Canal
IoU	Intersection over Union
JPG	Joint Photographic Experts Group
JSON	JavaScript Object Notation
LLM	Large Language Model
MAT	*MATLAB* file
MHA/MHD	*MetaImage* file
MRI	Magnetic Resonance Imaging
NIfTI	Neuroimaging Informatics Technology Initiative
OCT	Optical Coherence Tomography
OOM	Out of Memory
PET	Positron Emission Tomography
PNG	Portable Networks Graphics
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSMA	Prostate-Specific Membrane Antigen
RTSTRUCT	Radiotherapy Structure Set
SAM	Segment Anything Model
ViT	Vision Transformer

Part I

BACKGROUND

1

INTRODUCTION AND MOTIVATION

The main goal of this thesis is to standardize the field of interactive medical image segmentation by establishing a foundation for the systematic development and evaluation of future methods. To this end, we address six core areas that reflect both the fundamental components of interactive segmentation and the broader need for community-wide standardization: (1) the *representation of human interactions*, including how different types of interactions can be defined and encoded; (2) *efficient inference and training strategies* that ensure segmentation models remain responsive to user input; (3) the *realistic simulation of interactions*, enabling reliable evaluation without requiring extensive human annotation; (4) a *taxonomy* for classifying interactive segmentation methods; (5) a *community-driven initiative* to co-organize interactive segmentation challenges with international collaborators; and (6) a *large-scale standardized benchmark dataset* that unifies 166 open-licensed datasets across nine imaging modalities and provides consistent simulation and evaluation protocols. A high-level summary of these six contributions is presented in Figure 1. Together, they aim to move interactive medical image segmentation toward a more organized, transparent, and collaborative future.

1.1 INTERACTIVE MEDICAL IMAGE SEGMENTATION AND ANNOTATION

Semantic segmentation is a cornerstone of medical image analysis, enabling the quantification of imaging biomarkers, such as tumor size [186–188], and the detection of overlooked abnormalities [189, 190]. In some cases, segmentation models even outperform clinicians, producing more stable and detailed predictions [191, 192]. Beyond delineating structures, segmentation is used as an auxiliary task for a wide range of other goals, including automated report generation [193, 194], survival prediction [195–197], and radiotherapy planning [198–200]. Despite its wide adoption, segmentation is highly dependent on large amounts of high-quality, densely annotated images [201–203]. Creating such datasets is a major bottleneck, as annotation requires specialized medical expertise and can be extremely time-consuming, especially for data like volumetric computed tomography (CT) or magnetic resonance imaging (MRI) scans [204–206] and gigapixel whole-slide images [207, 208].

Interactive segmentation mitigates the dependency on manual annotation by offering an efficient shortcut for data labeling [149, 209]. It is a human-in-the-loop process where users iteratively provide corrective inputs, such as clicks, scribbles, or bounding boxes, to guide and refine the model's predictions. Interactive segmentation has demonstrated remarkable potential, with some studies reporting reductions in annotation time of more than 500-fold

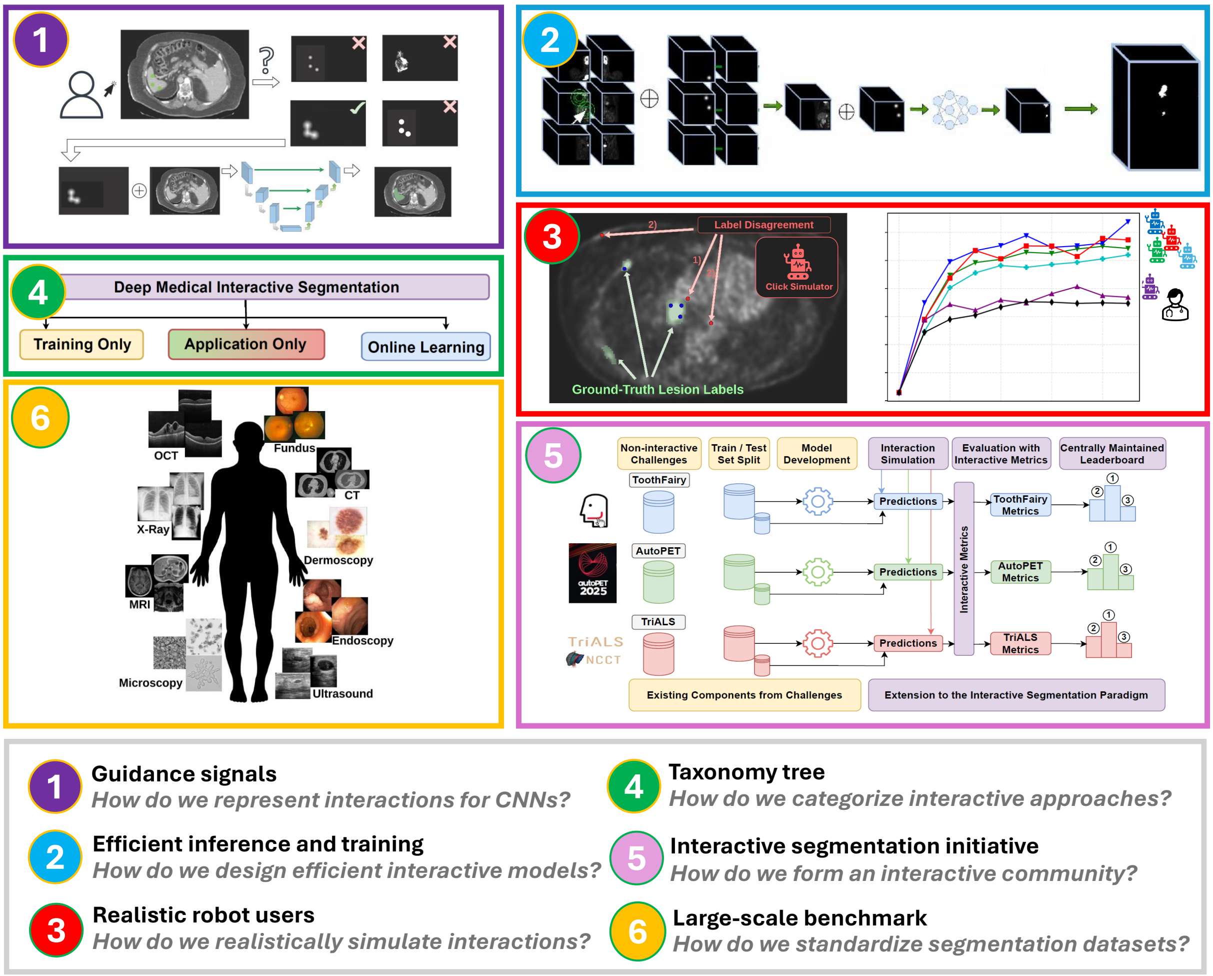

1
2
3
2)
Label Disagreement
1)
Click Simulator
Ground-Truth Lesion Labels
4
Deep Medical Interactive Segmentation
Training Only
Application Only
Online Learning
6
OCT
Fundus
CT
X-Ray
Dermoscopy
MRI
Endoscopy
Microscopy
Ultrasound
5
Non-interactive Challenges
Train / Test Set Split
Model Development
Interaction Simulation
Evaluation with Interactive Metrics
Centrally Maintained Leaderboard
ToothFairy
AutoPET
TriALS
autoPET 2025
TriALS NCCT
Predictions
Interactive Metrics
ToothFairy Metrics
AutoPET Metrics
TriALS Metrics
Existing Components from Challenges
Extension to the Interactive Segmentation Paradigm
1 Guidance signals
How do we represent interactions for CNNs?
2 Efficient inference and training
How do we design efficient interactive models?
3 Realistic robot users
How do we realistically simulate interactions?
4 Taxonomy tree
How do we categorize interactive approaches?
5 Interactive segmentation initiative
How do we form an interactive community?
6 Large-scale benchmark
How do we standardize segmentation datasets?

Figure 1: Overview of the contribution areas and their corresponding research questions in this PhD thesis. We investigate core components of interactive segmentation, focusing on: (1) representations of interactions; (2) efficient inference and training strategies; and (3) realistic interaction simulation. We also address the standardization of the field through: (4) a taxonomy for classifying interactive methods; (5) a community-driven initiative to co-organize interactive challenges; and (6) a large-scale standardized benchmark dataset.

for CT multi-organ annotation [92], transforming manual labeling that can require over 90 minutes per CT image into just a few quick interactions [206]. Beyond improving labeling efficiency, interactive models also enhance prediction quality, as annotators actively inspect and refine results during interaction [35, 62, 77]. When combined with active learning [22, 66], where only the most informative samples are selected for annotation, interactive methods have enabled the creation of large-scale datasets comprising thousands of images across diverse domains such as radiology [92] and histopathology [26].

From both a methodological and ethical perspective, incorporating human interactions into medical artificial intelligence (AI) systems is essential. Methodologically, it provides models with richer information and guidance, leading to more accurate and robust predictions [149, 209]. Ethically, it reinforces the role of AI as an assistant rather than a replacement for clinicians, an important distinction that helps build trust among both doctors and patients [210].

1.2 ON THE DESIGN AND EVALUATION OF INTERACTIVE SEGMENTATION MODELS

While all of these advancements may suggest that the problem is solved and interactive models have revolutionized medical AI, a simple question arises when one considers adopting interactive models in their workflow:

What is the state-of-the-art interactive segmentation model for my task?

Given the high relevance of this field, with segmentation models accounting for more than 20% of published papers[1] at *MICCAI*[2] in 2025, it is reasonable to expect a straightforward answer to this question. However, a concerning trend has emerged over the past nine years: more than 50% of deep learning-based interactive models published between 2016 and 2023 do not compare their results to any prior work [209]. Among those that do, comparisons are often arbitrary and lack a consistent pattern. Furthermore, methods employ a wide variety of metrics and datasets, making it difficult to obtain a clear overview of the field. Thus, this simple question becomes impossible to answer.

This disorder can largely be attributed to the absence of a central benchmark for medical interactive methods, which prevents researchers from comparing to each other's work. The gap cannot be attributed to the **interactive** aspect, as well-defined interactive benchmarks exist in non-medical domains and are paired with established evaluation metrics and interaction simulation protocols, including *DAVIS* [142], *GrabCut* [141], *Berkeley* [211], and *SBD* [144]. Nor is it caused by the **medical** domain itself, which already features widely adopted segmentation competitions[3] and open benchmarks, including *The Medical Segmentation Decathlon* [163] and the *TouchStone* benchmark [212].

Apart from the lack of a standardized framework for comparing interactive models, there is also no consensus on which design decisions are most effective for interactive models in the medical domain [149, 209, 213]. By contrast, the non-medical interactive domain features extensive analyses on how to represent user interactions [214, 216–219], as well as a consensus on how to realistically simulate interactions for benchmarks [214, 215, 220]. Furthermore, there are architecture-specific guidelines for adapting non-interactive convolutional neural network (CNN) models to effectively incorporate interactions [216, 221]. In contrast, the medical interactive segmentation field lacks both consensus on these topics as well as formal definitions of

[1] https://papers.miccai.org/miccai-2025/

[2] One of the largest conferences for medical AI.

[3] https://miccai.org/index.php/special-interest-groups/challenges/miccai-registered-challenges/

the fundamental concepts, making it difficult to even initiate a discussion in the community [149, 209, 213].

1.3 THESIS ROADMAP AND CONTRIBUTIONS

This thesis addresses the lack of standardization in interactive medical segmentation in four main areas:

1. **Theoretical foundations:** Formal definitions of key components, including interaction representation, interaction simulation, and a taxonomy for classifying interactive models, all based on our publication in *TPAMI 2024* [209].

2. **Practical recommendations:** Flowchart-based guidelines for selecting interactive models for specific applications (as published in *TPAMI 2024* [209]), together with evidence-based best practices for interaction representation (published in *MICCAI 2023* [91]), realistic interaction simulation (accepted in *MICCAIw 2024* [285]), and efficient interactive model inference and training (based on our publication in *ISBI 2024* [239]).

3. **Community engagement:** An international collaboration with four other research groups on co-organizing interactive segmentation competitions at *MICCAI 2025* and *CVPR 2025*, defining interactive segmentation tasks to enable participants to compete in developing the best solutions.

4. **Standardized large-scale dataset:** Consolidation of 166 public datasets across nine imaging modalities with standardized interaction simulation and evaluation protocols.

Our contributions establish theoretical foundations for interactive models, provide practical guidelines for their design and evaluation, engage the community through competitions, and offer a standardized dataset to ensure reproducibility and comparability across methods.

We organize our contributions into two parts in the thesis. In Part ii, we focus on the fundamental components of interactive segmentation: interaction representation, simulation, and efficient responsiveness to interactions. In Part iii, we address our standardization of the research landscape.

Specifically, in Chapter 3, we introduce a formal definition of the interaction representation (“guidance signal”), conduct a comparative study of existing guidance signals, and outline best practices for choosing a guidance signal. Chapter 4 builds on this by exploring methods to enhance model responsiveness to user interactions through more efficient inference and training. Finally, Part ii concludes with Chapter 5, which defines interactive simulation (“robot user”) and analyzes realistic user interaction simulation for whole-body lesion segmentation.

In Part iii, Chapter 6 provides a systematic review of existing deep learning-based interactive models, constructs a taxonomy to categorize them, and highlights pitfalls and research gaps in the field. Chapter 7 describes our co-organization of segmentation challenges through the *Interactive Segmentation Initiative*, engaging the community to collaboratively advance the field. Finally, in Chapter 8, we introduce *OmniMedSeg*, a standardized, large-scale multimodal dataset complete with interaction simulation and evaluation protocols.

A NOTE ON IMPLEMENTATION Zdravko Marinov was responsible for the implementation of the frameworks and experiments presented in *all sections* of Chapter 3, Chapter 5, Chapter 7, and Chapter 8. In Chapter 4, *all sections* are implemented solely by Zdravko Marinov, with the exception of Section 4.1, which is based on joint work carried out in close collaboration with his Master's student, Matthias Hadlich. Both Zdravko Marinov and Matthias Hadlich contributed substantially to this research. Zdravko Marinov conceptualized the ideas, designed the frameworks, and defined the experiments, while his student focused primarily on the code implementation. In Chapter 6, *all sections* were conceptualized in collaboration with Paul Jäger. Zdravko Marinov wrote the manuscript, collected and reviewed all studies, summarized the results, constructed the taxonomy tree, and described the observed trends among the reviewed works. Paul Jäger contributed by providing valuable feedback and insights during discussions.

2

RELATED WORK

This thesis draws on multiple research subfields in medical interactive segmentation, spanning both theoretical studies of interactive segmentation models and practical annotation tools used in clinical settings. While our results contribute to several areas, they are unified by the overarching theme of establishing common standards that can be applied to any interactive model, both in design and evaluation. This chapter provides an overview of the most relevant related work and highlights how our contributions extend and impact each area.

2.1 REPRESENTATIONS OF HUMAN INTERACTIONS

Human interaction is the defining element that sets interactive segmentation models apart from conventional semantic segmentation approaches. Prior research has proposed a wide range of strategies for encoding human interactions, enabling deep learning models to incorporate them either as additional input channels or as auxiliary components within the loss function or post-processing pipeline. These interactions take many forms, such as clicks [67, 214, 217], scribbles [43, 70], bounding boxes [1, 141], text [203, 227, 231], lassos [230, 268], and eye gaze [228, 229], among others [175, 209, 232, 233]. Despite their diversity, all these interaction types share a common requirement: they must be represented in a form that can be effectively processed by a semantic segmentation model.

For instance, in click-based interactions, common representations include Gaussian heatmaps [40, 67, 80, 222, 223], Euclidean distance transforms [140, 214, 215, 289], geodesic distance transforms [5, 38, 54], disc encodings [11, 221, 224], and positional encodings designed for transformer-based architectures [137, 201, 202]. For scribble-based interactions, representations typically include Euclidean [13, 53] and geodesic maps [22, 81], as well as parametric forms such as B-splines [43, 60, 70]. Bounding boxes are generally represented either implicitly by cropping the image to the box region [1, 9, 38, 40] or explicitly by encoding their corner points in a manner analogous to click-based inputs [35, 37, 51]. Text interactions are usually mapped to low-dimensional embeddings and integrated multimodally with the image [203, 227, 231], and eye gaze inputs are often treated as gaze-based scribbles and represented in a similar way to hand-drawn scribbles [228, 229].

Beyond input representations, user interactions can also be incorporated directly into the loss function during training. This is often achieved by assigning higher penalties to prediction errors near interaction points [65, 78, 85]. Interactions can also be utilized during post-processing to remove predictions that contradict user input [56, 58, 60]. For instance, lassos are frequently integrated as active contour post-processing [230, 268], refining the segmentation boundary toward local gradient minima in the image.

This diversity of interaction representations raises the question of which encoding type is most effective for interactive deep learning models. Previous studies have reported mixed results: some suggest that disc encodings provide the best performance for click-based inputs [221], while others highlight the advantages of geodesic maps [38, 81, 225] or Gaussian heatmaps [176, 226]. However, most of these conclusions stem from small-scale ablation studies in which representation analysis was not the primary focus. The only large-scale comparison to date [214], conducted on natural images, demonstrated that disc representations significantly outperformed alternative methods for click-based interactions. Furthermore, the broad variety of interaction types and their corresponding representations calls for a standardized framework that unifies these representations under a single, consistent definition—something that remains absent in existing literature.

Our contribution: Building on the insights from the representation study conducted on natural images [214] and given the wide range of existing interaction representations, our goal is to bring structure to this diversity. In Section 3.1, we formally define the main categories of interactive representations, collectively referred to as *guidance signals*, to help researchers identify and relate existing approaches. Section 3.2 presents our comparative analysis of click-based guidance signals evaluated on two medical segmentation tasks, highlighting the strengths and limitations of each signal type. The insights gained from this analysis motivate us to develop a novel hybrid guidance signal that combines the locality of Gaussian heatmaps with the global context of geodesic representations.

Impact: Our contributions presented in Chapter 3 demonstrate that interaction representations warrant dedicated investigation, as they are critical components of interactive models and should not remain confined to small ablation studies. Furthermore, we establish a theoretical framework that introduces consistent definitions for the various representations, enabling future work to build upon a shared conceptual foundation and terminology.

2.2 REALISTIC SIMULATION OF MEDICAL ANNOTATORS

Interactive models depend on human interactions for proper evaluation. In prior work, this is typically achieved in one of two ways: either through a user study with medical annotators using the interactive model [5, 65, 82], or by simulating the user interactions [51, 53, 81]. While studies with real annotators provide the most accurate assessment of practical usability, the vast majority of interactive models[1] (over 80%) rely on simulated interactions rather than user studies [209], likely due to the high cost of involving medical experts. The same pattern is observed during training: because interactive models require a large number of iterations, most models also simulate interactions during training [209].

[1] Deep learning-based medical interactive segmentation models published between 2016–2023

Similar to interaction representations, interaction simulations exhibit considerable diversity. Interactions can be simulated non-iteratively by generating all interactions at once and providing them to the model as additional input [13, 16, 44]. Alternatively, interactions can be simulated iteratively [2, 29, 74], where each subsequent interaction is determined based on the model's prediction error from previous interactions, mimicking how a real user would inspect and correct the model's output. Both iterative and non-iterative simulations can be either rule-based, where interactions are determined according to fixed rules [3, 12, 14], such as the object center or the location of the largest error, or sampling-based [2, 5, 55], where interactions are drawn from the ground-truth label and/or the model's previous prediction errors. This diversity makes it challenging to provide an overview of the current research in interaction simulation, and there is a lack of standardized definitions to categorize prior work.

While these simulation methods differ substantially and have their own specific characteristics, they all rely on a fundamental assumption: interactions must always conform to the ground-truth labels. Although this assumption seems intuitive, and simulating "incorrect" interactions might appear counterproductive, some tasks are known to exhibit low inter-annotator agreement. For instance, inter-annotator agreement measured by intersection over union (IoU) has been reported to be approximately 60% for cervical lesions in colposcopy [234] and around 54% for tuberculosis in chest X-rays [234–236]. In contrast, higher agreement has been observed for anatomical structures, with IoU values of 93% for the retina and 81% for the choroid in optical coherence tomography (OCT) images [237], and 85% for the gallbladder, 90% for the stomach, and 81% for the pancreas in CT images [92]. When interactions are simulated from these ground-truth labels, one can expect a similar level of variability to that observed among human annotators [285]. This raises an important question: How can interactions be realistically simulated given the inherent annotation noise?

Our contribution: In Chapter 5, we make two main contributions to interaction simulation. First, in Section 5.1, we formally define simulation methods, termed *robot users*, and categorize prior approaches. Second, in Section 5.2, we compare prevalent click-based simulation methods for positron emission tomography (PET)/CT whole-body lesion segmentation and find that existing simulations often produce unreliable metrics compared to a real user study with medical annotators using the same data and model. Motivated by this, we design a new simulation method that incorporates task-specific inter-annotator disagreement, simulating clicks outside the ground-truth to produce more realistic performance estimates. We validate this approach in two independent user studies to show our findings can be consistently reproduced.
Impact: Our definitions of robot users establish a standardized theoretical framework, enabling a more structured approach to designing and selecting interaction simulation methods. Our results also highlight that current simulation methods rely on legacy approaches that poorly reflect annotator behavior and ignore task-specific factors. By providing a novel simulation method for PET/CT lesions, we show that it is possible to approximate real-world performance accurately through thoughtful, task-informed simulation rather than directly applying existing simulation methods, a finding that can be applied beyond the PET/CT domain.

2.3 EFFICIENT INTERACTIVE MODELS

In clinical settings, interactive models must meet key criteria to be practically useful: they should be responsive and deliver predictions with low latency to enable seamless integration into clinical workflows. Together, these aspects determine a model's overall *efficiency*. Prior work has shown that efficiency is a crucial factor in interactive systems, as highlighted in user studies and questionnaires [19, 43, 85, 147, 148, 238]. Efficiency of interactive models can be assessed through metrics such as time measurements [5, 13, 38, 56] or the number of interactions required to reach a target prediction quality [51, 64, 72, 74].

However, most user studies focus narrowly on evaluating the efficiency of specific models [5, 38, 64, 74] or annotation tools [70, 149], often justifying the choice of a particular system rather than analyzing which components most influence a model's efficiency and responsiveness. In contrast, we investigate how fundamental elements of interactive systems shape overall model efficiency, deriving practical recommendations from user studies conducted with medical annotators. Our findings are designed to generalize across interactive systems, offering insights that extend beyond the evaluation of individual models or tools.

Our contribution: In Chapter 4, we investigate high-level factors that influence the efficiency of interactive systems, focusing primarily on improving model inference and training. In Section 4.1, we introduce a strategy for optimizing sliding window–based models through *local corrective inference* and evaluate its impact on model efficiency through a user study involving medical annotators. Then, in Section 4.2, we present methods to reduce the training time of interactive models by optimizing the computation of distance transforms for interaction simulation during training.
Impact: Our findings show that model efficiency extends beyond evaluating specific models; it can inform best practices for model-agnostic components such as inference mechanisms and training strategies. These insights have broad applicability across interactive segmentation and annotation systems, offering a foundation for the development of more efficient clinical tools.

2.4 REVIEW AND TAXONOMY OF DEEP MEDICAL INTERACTIVE SEGMENTATION

The field of deep learning-based medical interactive segmentation is expanding rapidly, with over 120 publications between 2016 and 2023 and the number continuing the rise every year [209]. As keeping up with this pace becomes infeasible, review papers and taxonomies are essential for summarizing developments, classifying approaches, and guiding researchers in finding relevant work. Several reviews on interactive segmentation exist. However, most either focus on classical approaches rather than the more recent deep learning methods [145, 155, 156, 240], or exclude approaches from the medical domain altogether [157]. At the same time, no review exists for the field of deep learning-based interactive segmentation of medical images despite its rapid emergence. Additionally, the only existing taxonomy for medical interactive segmentation methods focuses on classical non-deep learning approaches and is already over a decade old [155].

Our contribution: In Chapter 6, we fill this long-standing gap by conducting a systematic review of deep learning-based medical interactive segmentation methods, covering 121 publications in the time period 2016-2023. In Section 6.1, we introduce a new taxonomy tree that organizes existing approaches into distinct categories. We design the taxonomy tree by asking a question at each junction and helping readers traverse the tree down to its terminal taxonomy categories by simply answering a series of questions. We also make practical recommendations on what category of interactive model to select, given a specific application. In Section 6.2, we discuss the concerning pitfalls and research gaps we have identified during our review, and in Section 6.3, we discuss potential opportunities on how to fill these gaps.
Impact: Our review and taxonomy provide a standardized structure to shape the entire field of deep medical interactive segmentation. They enable researchers to systematically classify existing methods, select suitable models for specific applications, and compare new approaches with prior work. Our review also identified major gaps in the field that have directly motivated several of the other contributions presented in this thesis.

2.5 THE MEDICAL INTERACTIVE SEGMENTATION COMMUNITY

The medical interactive segmentation community has grown in recent years, driven by a collective shift toward openness, reproducibility, and collaboration. One of the most established and organized projects in this space is *MONAI Label* [149], which functions as both a framework[2] and a community hub. It is fully open-source, hosts regular workshops to engage with its users, and provides templates for training and deploying certain interactive segmentation models such as *DeepEdit* [67] and *NuClick* [26]. *3D Slicer* [241] represents another major pillar of the community, with an extensive global user base, active discussion forums[3], and a strong ecosystem of extensions that facilitate data annotation and visualization. Other open-source frameworks, including *VesselVerse* [245], *BioMedisa* [151], *AnatomySketch* [70], *RIL-contour* [150], *MITK* [173], *Vessel-CAPTCHA* [232], and *PyMIC* [174], further enrich the landscape by supporting interactive and semi-automatic segmentation workflows across diverse imaging modalities. Complementary to these efforts, non–deep learning frameworks such as *ilastik* [168] and *ITK-SNAP* [169], continue to play an important role in data annotation. Another important aspect that helps the community grow is the increasing availability of open datasets and competitions on platforms such as *Kaggle*[4], *Grand Challenge*[5], and *Synapse*[6], which strengthens collaboration and benchmarking within the segmentation community.

[2] https://github.com/Project-MONAI/MONAILabel
[3] https://discourse.slicer.org/
[4] https://www.kaggle.com
[5] https://grand-challenge.org/challenges/
[6] https://www.synapse.org/

Although open-source projects foster community collaboration to advance segmentation methods, the main focus is shifted primarily on providing high-quality software tools that allow researchers to implement their own approaches [149, 169, 174] or are focused on a very specific annotation task [151, 232, 245] rather than addressing best practices for improving the fundamental components of interactive systems, such as interaction representation, realistic interaction simulation, model categorization, or identification of research gaps. In contrast, segmentation competitions, also termed *challenges*, promote the discovery of best practices through collective experimentation. Their competitive structure encourages the emergence of novel methods [146] and enables fair, consistent evaluation of all submissions due to their centralized evaluation [246]. Challenges also cultivate sub-communities of organizers and participants, who often collaborate on post-challenge publications to share findings with the community and potentially build upon them in future iterations of the challenge. Additionally, challenges also host workshop sessions when presenting the final results, which gives a platform for active discussion between everyone involved.

However, interactive segmentation challenges remain rare, with the only exception being the *CVPR 2024: Segment Anything in Medical Images on Laptop* challenge [242] — the first interactive challenge in over 13 years following the *PROMISE12* competition [243]. Inspired by the success of this challenge [242], we aim to expand its efforts to a multi-challenge initiative by joining forces with its organizers and three other long-standing challenges at *MICCAI* (*autoPET* [154], *ToothFairy* [244], and *TriALS*[7]), to focus specifically on interactive segmentation.

Our contribution: In Chapter 7, we describe how we establish a global interactive segmentation community by co-organizing four international competitions across multiple centers, attracting over 200 participant submissions. In Section 7.2, we explain how the community is formed in collaboration with other challenge organizers by extending existing challenges with an interactive layer. In Section 7.3, we present how we unify all challenges under standardized metrics, interaction types, and baseline models, enabling the identification of challenge-agnostic trends that reveal what makes an interactive model perform well across tasks. Finally, in Section 7.4, we analyze the key components contributing to the success of the winning models.

Impact: Our interactive segmentation initiative brings research gaps in the field to the forefront at a community level. By providing a framework for successfully hosting interactive segmentation challenges, we enable other competitions to join the initiative, fostering exploration of additional research dimensions and a broader range of segmentation tasks.

2.6 MEDICAL IMAGE SEGMENTATION DATASETS

Medical imaging datasets are highly diverse, spanning multiple modalities in both 2D and 3D. Common 2D modalities include fundus photography [247, 248], X-ray [249, 250], ultrasound

[7] https://www.synapse.org/Synapse:syn65878273/wiki/631556

[251, 253], endoscopy [252, 254], OCT [237, 255], microscopy [256, 257], and dermoscopy [258, 259], while 3D modalities include CT [92, 260], MRI [243, 261], and PET [154]. Each modality can be further subdivided into subtypes, such as different microscope types for microscopy, T1, T2, or fluid-attenuated inversion recovery (FLAIR) sequences for MRI, and various PET tracers, e.g., fluorodeoxyglucose (FDG) or prostate-specific membrane antigen (PSMA), contributing to significant variability in imaging data.

Datasets also vary widely in file formats. For 2D images, common formats include *Portable Network Graphics* (PNG) and *Joint Photographic Experts Group* (JPG), though legacy formats such as *MATLAB* files (MAT) are still in use. For 3D volumes, standard formats include *Neuroimaging Informatics Technology Initiative* (NIfTI) files, *Digital Imaging and Communications in Medicine* (DICOM) series, *MetaImage* (MHA/MHD), and *Hierarchical Data Format 5* (H5). Labels exhibit similar diversity: in 2D, they may appear as simple binary masks or text files encoding contours, while in 3D, labels are often stored as binary NIfTI files, DICOM-Segmentation (DICOM-SEG) objects, or radiotherapy structure sets (RTSTRUCT). More unconventional approaches also exist, such as markers overlaid on images via tablets or compact representations like run-length encoding for space efficiency.

The main takeaway from this diversity is that medical imaging datasets are so heterogeneous that significant effort is required to combine them consistently for specific tasks. Prior work has sought to standardize data by pre-processing it into a uniform format before releasing it to the community. For example, Ma et al. [201] pre-processed over one million medical image–mask pairs by resizing images to 1024 × 1024, normalizing CT images using a fixed window for specific organ types and normalizing other modalities via percentile clipping, finally saving the results as *NPZ* files. However, such preprocessing discards a large portion of the information in the raw data, such as Hounsfield Units in CT scans or metadata in DICOM files, which could be valuable for downstream use. Additionally, because the dataset was prepared for a bounding box-based model *MedSAM* [201], images containing only objects smaller than 100 pixels in 2D or 1000 voxels in 3D[8] were omitted, altering the dataset's integrity. This also removes many axial slices in volumetric data where the object is absent or too small, effectively discarding global context that models could potentially leverage during training and influencing the final evaluation metrics. Other pre-processed datasets have similarly omitted small objects and metadata [213, 262], following the practices proposed by Ma et al. [201].

While we agree that standardization is crucial, particularly for applying standards across multiple datasets (e.g., simulating clicks for interactive segmentation), it must be implemented in a generic and flexible way. Standardized, strongly filtered datasets are convenient and easy to use, but results obtained from them are inherently biased toward large objects and do not reflect performance on unfiltered clinical data. Moreover, discarding parts of the datasets limits the applicability of these datasets, forcing models to learn biases encoded in the data preprocessing and making it difficult to interpret evaluation metrics. Similarly, normalization

[8] All axial slices that only cover connected components of objects with a size below 1000 voxels.

procedures and the removal of metadata restrict the clinical utility of the data and reduce the user's freedom to select and utilize information according to their needs.

Similar to prior efforts, we aim to standardize a large-scale, multi-modal segmentation dataset by utilizing public data. However, unlike previous work [201, 213, 262], our goal is to preserve the original raw data and labels as much as possible in the dataset's final form, even providing users with access to the original data. We also focus exclusively on directly accessible, open-licensed datasets to allow users to have as much freedom as possible with the use of the data. By standardizing datasets and labels to a consistent input format, we can apply our generic findings from the rest of this thesis to establish standardized click simulation protocols and evaluation metrics across all datasets, providing a platform to develop and evaluate interactive segmentation models in a standardized way.

Our contribution: In Section 8.2, we present *OmniMedSeg* — a large-scale, multimodal dataset that consolidates 166 open-licensed datasets spanning nine imaging modalities. We describe the data collection process, conversion to a standardized format, and the definition of unified protocols for interaction simulation and evaluation, promoting transparent and reproducible benchmarking across medical imaging domains. The dataset preserves all original metadata and raw imaging data, traceable to their sources, and employs a minimal preprocessing pipeline to maintain data integrity.
Impact: *OmniMedSeg* serves as a foundation for standardized training and evaluation of both non-interactive and interactive segmentation methods, fostering greater accessibility, reproducibility, and comparability in the medical image segmentation research.

Part II

FOUNDATIONS OF INTERACTIVE SEGMENTATION

3 INTERACTION REPRESENTATION

The story of this thesis begins at the input level of interactive segmentation models. Since interactions are abstract by nature, representing the actions users perform to influence the model, they must first be transformed into a suitable *representation* that the model can process and integrate into its prediction pipeline. In Section 3.1 (based on our *TPAMI 2024* publication [209]), we formally define this representation and introduce the concept of a *guidance signal*. We then discuss the various categories of guidance signals and how they can be integrated into deep learning models. Section 3.2, based on our *MICCAI 2023* publication [91], presents a comparative analysis of the most prevalent click-based guidance signals for two medical imaging tasks. This study reveals the fundamental strengths and limitations of existing signals and motivates the design of a novel hybrid representation that combines Gaussian heatmaps with geodesic distances, effectively addressing the key shortcomings of both approaches.

3.1 WHAT IS A GUIDANCE SIGNAL?

This section is based on parts of our publication in *TPAMI 2024* [209], © IEEE.

3.1.1 *Interaction Types*

We first define the interaction types used for deep learning-based interactive segmentation in prior work over the last decade [209] - clicks, scribbles, bounding boxes, and polygon vertices. These are illustrated in Figure 2.

Clicks are the most frequent interaction type, employed by over 50% of deep learning-based segmentation models [209]. A click c is defined as a 2D or 3D point, i.e., $c \in \mathbb{N}^2$ or $c \in \mathbb{N}^3$, depending on the dimensionality of the input image.

Scribbles are used by 35% of interactive approaches [209] and are also referred to as brush strokes [122, 266] or freehand sketches [267, 268]. Formally, a scribble $\mathcal{S}$ is a set of 2D or 3D points:

$$\mathcal{S} = \{p_1, \ldots, p_N\}, \quad p_i \in \mathbb{N}^D, \quad N \in \mathbb{N}, \quad D \in \{2, 3\} \tag{1}$$

where N is the number of points. In a formal sense, scribbles can be considered as a set of clicks; in practice, they represent a diverse range of interactions, including structured line strokes [36] and boundary contours [27], as well as unstructured marks such as random dabs

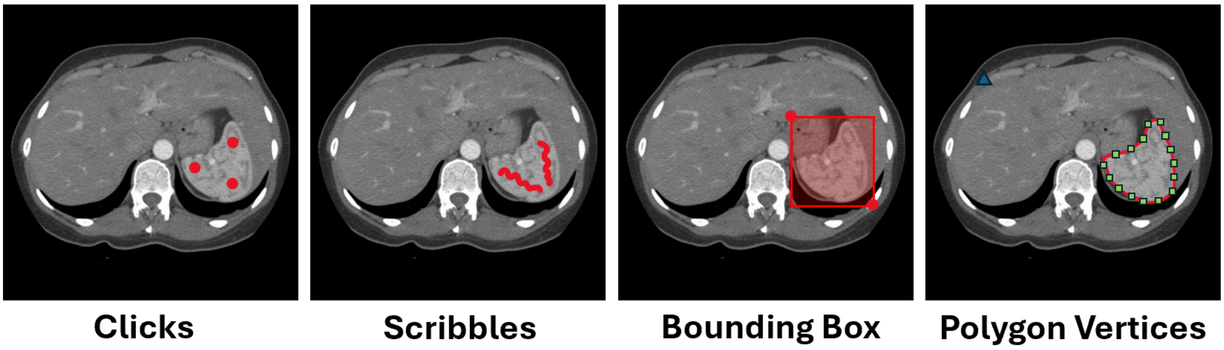


Figure 2: The four main interaction types used in prior work for deep medical interactive segmentation.

[5], or any combination of these [209]. Therefore, we adopt a broad definition to capture all these variations.

Bounding Boxes are another common interaction type, used in 24% of deep interactive models [209]. Bounding boxes indicate a region of interest containing the segmentation target. Formally, a bounding box $\mathcal{B}$ is the set of coordinates enclosed by two opposing corner points b_{min} and b_{max}:

$$\mathcal{B} = \{p \in \mathbb{N}^D \mid b_{\text{min}}^{(d)} \leqslant p^{(d)} \leqslant b_{\text{max}}^{(d)},\ \forall\, d \in \{1, \ldots, D\}\}, \quad D \in \{2, 3\}. \tag{2}$$

Here, b_{min} and b_{max} correspond to the top-left and bottom-right corners (or analogous 3D coordinates), and $x^{(d)}$ is the d-th element of a vector x. Bounding boxes can be interpreted as a structured set of points with a predefined rectangular or cuboidal topology.

Polygon Vertices are a rarer interaction type, used by roughly 7% of prior approaches [209], in which a user specifies the vertices of a boundary polygon to approximate the true object boundary. Formally, we represent this as:

$$\mathcal{P} = \{p_1, \ldots, p_N\}, \quad p_i \in \mathbb{N}^2, \tag{3}$$

where N is the number of polygon vertices. Prior work has applied polygons exclusively in 2D. Polygon vertices can be considered as a set of ordered points on the boundary of the object.

Other interaction types include eye gaze [228, 229], which prior work typically reduces to scribbles by recording the set of coordinates observed by the user, and text prompts [203, 231, 269]. Text-based interactions for medical image segmentation, however, have gained traction only after our systematic review [209] and are therefore outside the scope of this thesis. Nevertheless, we discuss their potential relevance for the future of the field in Section 10.2.

3.1.2 *Definition of a Guidance Signal*

Interaction representations are often termed as *guidance signals* in literature [10, 149, 219, 263–265]. We believe this term reflects precisely the role of the representation in the interactive model - it guides the model towards the desired behaviour and provides a signal as additional information to the image. Hence, in our *TPAMI 2024* publication [209], we use this term to define the interactive representation and use this terminology throughout the thesis. The exact definition of a guidance signal is as follows:

Definition 1: A *guidance signal* encodes user interactions in a form that the model can process. It can be **explicit**, where interactions are transformed into structured inputs, such as Gaussian heatmaps centered on user clicks, or **implicit**, where interactions are incorporated into the model's training or inference without using a defined structure, for instance by weighting the loss function according to the distance to user clicks, or cropping the image based on a bounding box. Implicit signals are conveyed through *actions*, whereas explicit signals represent *structured entities*.

3.1.3 *Explicit Guidance Signals*

The interaction types introduced in Section 3.1 are defined as sets of discrete coordinates. Explicit guidance signals map these coordinates onto new structured entities. Since these entities have a defined structure, we can directly visualize examples in Figures 3 and 4.

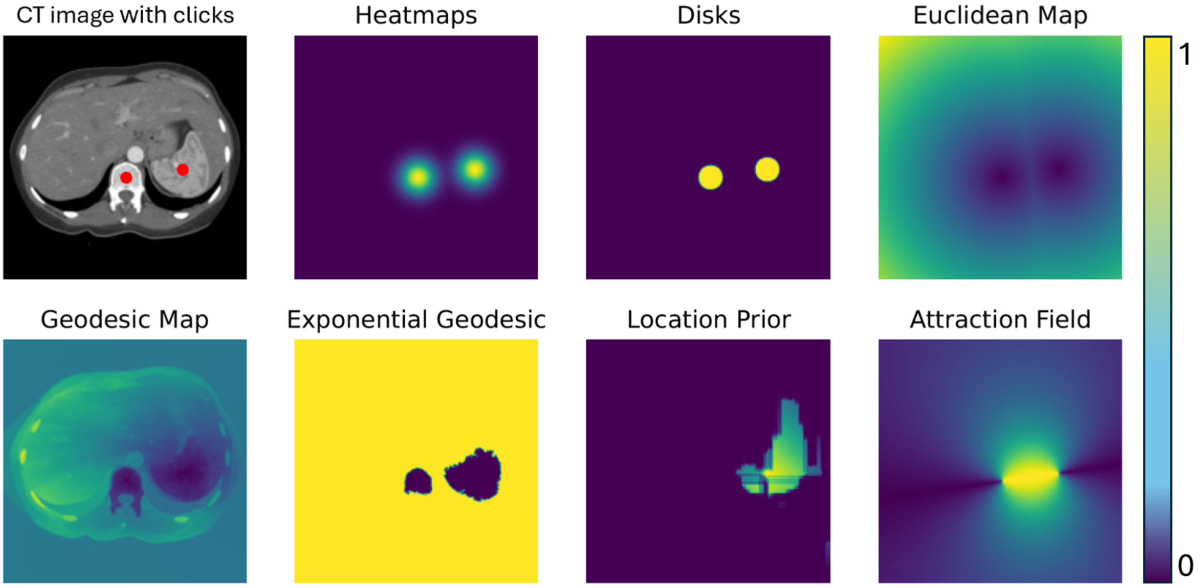


Figure 3: Examples of explicit guidance signals for **clicks** used in the majority of prior work.

The guidance signals shown in Figures 3 and 4 preserve the same spatial resolution as the original image. Consequently, many models integrate these explicit signals via simple concatenation with the input image in an early-fusion manner [10, 38, 140, 209], relying on the network to implicitly learn how to exploit the interaction information. Explicit guidance sig-

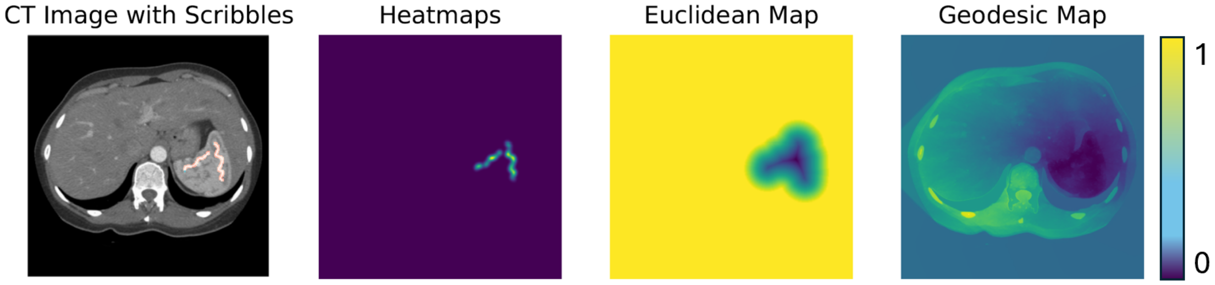


Figure 4: Examples of explicit guidance signals for **scribbles** used in the majority of prior work.

nals also offer a degree of interpretability, as users can visually verify whether they reflect their intended interactions. Moreover, certain signals, such as geodesic maps, encode a coarse pre-segmentation of the target object, since their computation partially depends on the underlying image features. Due to their intuitive and tangible nature, explicit guidance signals are adopted by 87% of all deep medical interactive models [209]. Here, we formally define the explicit guidance signals that we compare in Section 3.2, while the remaining explicit guidance signals we identified in our *TPAMI 2024* [209] work, are provided in the Appendix Section A.1.

Click-based Explicit Guidance Signals are visualized in Figure 3. Let the clicks provided by the annotater be: $\mathcal{C} = \{c_1, ...c_N\}$, where N is the number of clicks. We define all guidance signals over the voxel/pixel coordinates $v \in \mathbb{N}^D$ of the image I, where $D \in \{2, 3\}$ for 2D images $I \in \mathbb{R}^{W \times H}$ or 3D images $I \in \mathbb{R}^{W \times H \times K}$.

Heatmaps apply Gaussian filters centered around each click c_i with a radius $\sigma \in \mathbb{N}$ to create softer edges with an exponential decrease away from the click, and are defined in Equation 4:

$$\text{heatmap}(v, c_i, \sigma) = \exp\left(-\frac{\|v - c_i\|_2^2}{2\sigma^2}\right) \tag{4}$$

Disks are computed similarly as the Gaussian heatmaps but are filled with a constant value instead of the exponential term in the heatmaps as seen in Equation 5:

$$\text{disk}(v, c_i, \sigma) = \begin{cases} 1, & \text{if } \|v - c_i\|_2^2 \leqslant \sigma \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

As disks and Gaussian heatmaps are computed independently for each click, they are defined for a single click c_i over voxels/pixels $v \in \mathbb{N}^D$ in the image I. When considering all of the annotator's clicks $\mathcal{C} = \{c_1, ...c_N\}$, Equations 4 and 5 are extended as:

$$\text{heatmap}(v, \mathcal{C}, \sigma) = \sum_{i=1}^{N} \text{heatmap}(v, c_i, \sigma), \quad \text{disk}(v, \mathcal{C}, \sigma) = \sum_{i=1}^{N} \text{disk}(v, c_i, \sigma) \tag{6}$$

Euclidean Distance Transforms (EDT), also called Euclidean maps, are defined in Equation 7 as the minimum Euclidean distance between a voxel/pixel v and the set of clicks $\mathcal{C}$. It is similar to the disk signal in Equation 5, but instead of filling the sphere with a constant value it computes the distance of each voxel/pixel to the closest click point.

$$\mathrm{EDT}(v, \mathcal{C}) = \min_{c \in \mathcal{C}} \|v - c\|_2^2 \tag{7}$$

Geodesic Distance Transforms (GDT), or geodesic maps, are defined in Equation 8 as the shortest geodesic path distance between each voxel/pixel v and the set of clicks $\mathcal{C}$ [270]. The shortest geodesic path in GDT also takes into account intensity changes between voxels/pixels along the path. The shortest geodesic path is denoted as Φ in Equation 8 and can be computed with, e.g., the Fast Marching method [271].

$$\mathrm{GDT}(v, \mathcal{C}) = \min_{c \in \mathcal{C}} \Phi(v, c) \tag{8}$$

The Exponentialized Geodesic distance proposed in *MIDeepSeg* [38] is defined in Equation 9 as an exponentiation of GDT from Equation 8:

$$\text{exp-GDT}(v, \mathcal{C}) = 1 - \exp(-\mathrm{GDT}(v, \mathcal{C})) \tag{9}$$

For scribble-, bounding box-, and polygon vertices-based explicit guidance signals, please refer to Appendix Section A.1.

3.1.4 *Implicit Guidance Signals*

Implicit guidance signals differ from explicit ones in that they encode *actions* rather than *structured entities*. Compared to explicit signals, they have been far less explored in prior work, appearing in only 13% of approaches [209]. In this section, we briefly discuss how user interactions can serve as implicit forms of guidance.

Click-based Implicit Guidance Signals have been employed in a variety of ways. Ju et al. [78] use clicks to enforce class consistency across all predicted voxels within specific appearance and distance constraints, similar to conditional random field-based post-processing [166]. Other works [56] leverage clicks to guide a superpixel-based method [133], propagating user-defined labels to neighboring superpixels according to their similarity in a pre-computed binary partition tree [132] with superpixels as terminal nodes. Clicks have also been used implicitly to condition a prior-generator network [34] to produce a coarse segmentation mask constrained within the convex hull of the provided clicks. Similarly, they have been employed as bounding box surrogates to define a fixed-size crop centered around the click location [28].

Scribble-based Implicit Guidance Signals have been utilized to enforce prediction constraints and to propagate class labels based on appearance and distance similarity [84], or alternatively, based on epistemic uncertainty [17]. Scribbles have also been incorporated as auxiliary components in loss functions, where predictions are weighted more strongly near user-provided interactions [18, 21, 31, 48, 65, 85].

Bounding Box-based Implicit Guidance Signals are typically implemented by cropping the image to the region of interest before forwarding it to the model [6, 8, 9, 21, 35, 40, 42, 51, 71, 72]. This simple yet effective strategy implicitly focuses the model's attention on the relevant spatial region.

Polygon Vertices-based Implicit Guidance Signals have been explored in only one known work [87], where polygon vertices are used to bilinearly interpolate feature vectors within the network toward the user-provided contour.

Overall, implicit guidance signals enforce the intent and behavior of the user rather than the explicit spatial structure of their interactions. While less common, they offer a powerful means of incorporating interaction dynamics into learning and inference.

3.2 A COMPARATIVE ANALYSIS OF GUIDANCE SIGNALS

This section is based on our publication in *MICCAI 2023* [91].

Having defined the various types of guidance signals used in the literature in Section 3.1, a natural question arises (as depicted in Figure 5):

How should one choose a guidance signal for their interactive model?

Answering this question is challenging due to the diversity of approaches and the limited number of systematic comparisons. Most prior work designs interactive segmentation systems without performing ablation studies, with only a few exceptions in the medical imaging domain [38, 81] and several in non-medical contexts [25, 176, 214, 221]. Kovács et al. [276] observe that "the literature presents a complex and somewhat inconsistent picture of the comparative performance of various user input encoding methods" and that "many works employ a specific encoding without thorough investigation of alternatives." Similarly, Dupont et al. [275] note that comprehensive comparisons of guidance signals are largely absent.

The closest related experiments are in *MIDeepSeg* [38], which introduce an exponentialized geodesic distance guidance signal and compare it to other approaches through ablation studies. However, their analysis is limited to a fixed set of initial user clicks and does not evaluate iterative corrective interactions, an essential aspect of interactive segmentation. Prior studies also generally apply guidance signals without exploring how to optimally tune the signal's hyperparameters [25, 81, 176], leaving open questions about how guidance signals should be fairly compared.

To address these gaps, we conduct a comparative study of the most prevalent click-based guidance signals across two medical segmentation tasks. We focus on explicit click-based signals, which are most commonly used in prior work [209], and restrict experiments to CNN-based architectures to isolate the effect of guidance signals from architectural differences. In our comparative analysis, we train interactive models with various signals and hyperparameter settings to identify factors that most strongly influence performance.

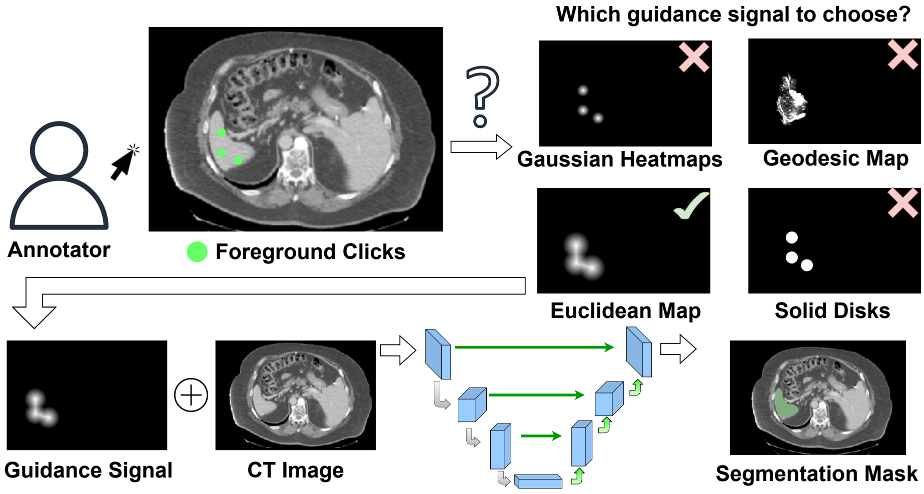


Figure 5: How should one choose a guidance signal when designing an interactive segmentation model? The large variety of guidance signals and the lack of best practices in selecting a signal make this simple question difficult to answer.

Building on these findings, we propose an adaptive Gaussian heatmap guidance signal that uses the geodesic distance transform to dynamically adjust the radius of each heatmap when encoding clicks. We evaluate this approach on the *MSD Spleen* [163] and *autoPET* [279] datasets, covering both anatomical (spleen) and pathological (tumor lesions) segmentation. Our results show that the choice of the guidance signal has a substantial impact on interactive segmentation, with our adaptive heatmaps achieving a 14% Dice improvement on the challenging *autoPET* dataset compared to non-interactive models.

Our study highlights the critical role of guidance signal selection and provides practical insights for designing more effective and efficient interactive segmentation systems. Our contributions can be summarized as follows:

1. We compare 5 existing guidance signals on the *autoPET* [279] and *MSD Spleen* [163] datasets and vary four guidance signal-specific hyperparameters **(H1)-(H4)**. We show which parameters are essential to tune for each guidance and suggest default values.

2. We introduce 5 guidance evaluation metrics **(M1)-(M5)**, which evaluate the performance, efficiency, and ability to improve with new clicks. This provides a systematic framework for comparing guidance signals in future research.

3. Based on our insights from 1., we propose novel adaptive Gaussian heatmaps, which use geodesic distance values around each click to set the radius of each heatmap. Our adaptive heatmaps mitigate the weaknesses of the 5 guidances and achieve the best performance on *autoPET* [279] and *MSD Spleen* [163] on 4 out of 5 metrics **(M1)-(M5)**.

3.2.1 *Experimental Setup*

3.2.1.1 *Interactive Segmentation Model and Datasets*

Guidance Signals. We consider explicit click-based guidance signals, which we formally defined in Section 3.1, focusing on five signal types: Gaussian heatmaps, Disks, Euclidean Distance Transforms (EDT), Geodesic Distance Transforms (GDT), and Exponentialized GDT, as defined in Equations 4–9. We exclude the other types of explicit guidance signals for clicks, defined in Appendix Section A.1, as they either focus only on 2D images (Location Priors), are constrained to exactly two clicks (Attraction Field Weight Maps), or focus on transformer-based architectures (Positional Encodings).

Model Architecture. We extend the *DeepEdit* model [67] from *MONAI Label* [149] to support these five click-based guidance signals. We focus specifically on *DeepEdit* as it is based on the U-Net architecture [127], which has been widely adopted by the medical image segmentation community [277], and it is the flagship interactive model of *MONAI Label* [67, 149]. The model consists of 6 downsampling convolutional layers with stride 2, follow by 6 upsampling deconvolutional layers with skip connections. Originally in *DeepEdit*, the click guidance signals, consisting of 2 channels for foreground and background clicks, are simply concatenated with the image as a joint input. However, we explore two other ways to integrate interactions into the model, which we discuss in Subsection 3.2.1.2.

Datasets. We conduct our comparative study using two publicly available datasets: *autoPET* [279] and *MSD Spleen* [163]. The *MSD Spleen* dataset contains 41 CT volumes with dense spleen annotations, a voxel spacing of $0.79 \times 0.79 \times 5.00$ mm^3, and an average resolution of $512 \times 512 \times 89$ voxels. The *autoPET* dataset comprises 1014 PET/CT volumes with annotated lesions from patients with melanoma, lung cancer, or lymphoma. We exclude the 513 lesion-free cases, resulting in 501 annotated volumes, and use only the PET modality for our experiments. The PET scans have a voxel spacing of $2.0 \times 2.0 \times 3.0$ mm^3 and an average resolution of $400 \times 400 \times 352$ voxels. The radioactive tracer used in *autoPET* is FDG, which highlights highly active metabolic regions in the body. Areas with high FDG uptake are commonly seen in the brain, heart, liver, and active tumors, as well as in some inflammatory or infectious processes [279].

Click Simulation. During both training and evaluation, we simulate a fixed number of clicks, N, following the procedure of Sakinis et al. [10]. For each volume, clicks are iteratively sampled from over- and under-segmented regions of the model's prediction by sampling points from the Euclidean distance transform of the largest connected error components. The sampled points are encoded as foreground and background guidance signals. The model performs a total of $N + 1$ prediction iterations, starting from zero clicks and incrementally adding one foreground and background click per iteration, resulting in 11 predictions when $N = 10$.

Model Hyperparameters. We keep all model hyperparameters consistent across experiments. Specifically, we use $N = 10$ clicks for each image, the Dice–Cross Entropy loss, a learning rate of 10^{-5}, and a fixed 80-20 train–validation split. Identical data augmentation strategies are applied to all models, consisting of random flips along each axis with a 10% probability and random 90° rotations with a 10% probability per axis. Models are trained on image crops with a fixed size of (192 × 192 × 256) using a single NVIDIA A100 GPU for 20 epochs on the *autoPET* dataset [279] and 100 epochs on the *MSD Spleen* dataset [163].

3.2.1.2 *Guidance Signal Hyperparameters*

We examine the impact of various hyperparameters specific to the guidance signals on the overall performance. This analysis enables us to provide recommendations for optimal parameter settings for each guidance signal. It also lets us compare the best-performing configurations of each guidance signal, ensuring a fairer and more meaningful analysis. Here, we define the four guidance signal hyperparameters **(H1) – (H4)**:

(H1) Sigma. We vary the radius σ of the disks and heatmaps defined in Equations 4 and 5. We also analyze how this parameter affects the performance of the distance-based signals, where instead of initializing the seed clicks $\mathcal{C}$ as individual voxels c, we define $\mathcal{C}$ as the set of all voxels within a radius σ centered at each c. The distance transform is then computed as in Equations 7, 8, and 9, meaning that the distances are calculated with respect to disks rather than single voxels.

(H2) Theta. We investigate how truncating the values of the distance-based signals defined in Equations 7–9 influences performance. Specifically, we discard the top $\theta \in \{0\%, 10\%, 30\%, 50\%\}$ of distance values, retaining only smaller distances closer to the clicks to provide more localized and precise guidance. Examples of how the θ hyperparameter influences the GDT computation are visualized in Figure 6. We do not explore this parameter on the exponentialized geodesic maps because the exponential term already suppresses large distances, and *MIDeepSeg* [38] proposes this guidance to exactly avoid using such thresholds.

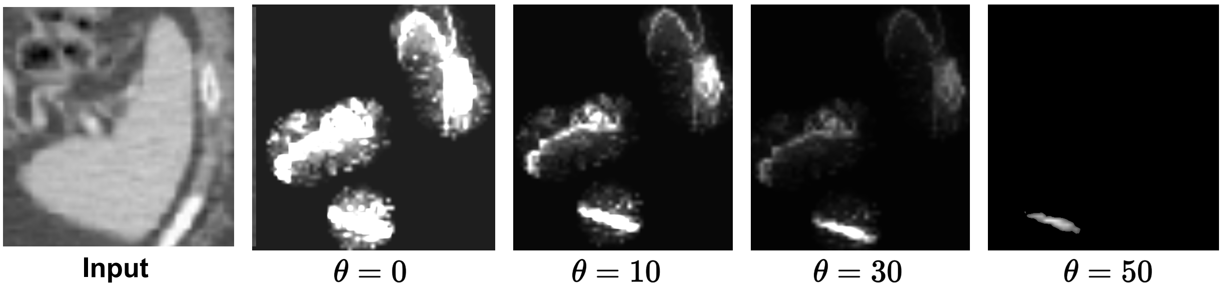


Figure 6: Example for varying the θ parameter **(H2)** in the GDT guidance signal. A low θ is noisy, but a high θ discards too many GDT values and leads to a weaker signal.

(H3) Input Adaptor. We evaluate three strategies for integrating guidance signals with input volumes, as proposed by Sofiiuk et al. [221]: Concatenation, Distance Maps Fusion (DMF), and Conv1S, depicted in Figure 7. In Concatenation, the image and guidance signals are combined by stacking them along the channel dimension. DMF further applies 1 × 1 convolutional layers to adjust the number of channels to match the backbone's original input size. Conv1S processes the image and guidance signals in separate branches, sums their feature representations, and then feeds the result into the backbone. DMF and Conv1S are designed to avoid changing the original backbone's parameters, e.g., by changing the number of input channels.

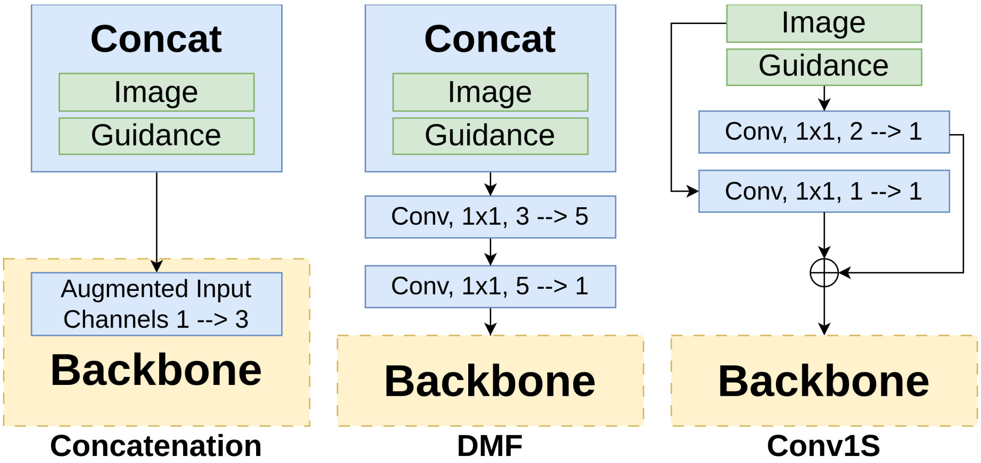


Figure 7: Visualization of the input adaptors **(H3)**. The input image contains a single channel, while the guidance signal consists of two channels representing foreground and background. The original backbone network is assumed to have an input layer configured for a single-channel input.

(H4) Probability of Interaction. As originally proposed for *DeepEdit* [67], during training, we randomly decide for each volume whether to provide the N clicks or not, with a probability of p, in order to make the model more independent of interactions and improve its initial segmentation. Thus, the expected value of interaction-free iterations during training is $(1 - p) \times |\mathcal{D}_{\text{train}}|$, where $\mathcal{D}_{\text{train}}$ is the training dataset.

All guidance-signal-specific hyperparameters that we vary are summarized in Table 1. Each combination of hyperparameters corresponds to a separately trained *DeepEdit* [67] model. In total, there are 495 possible hyperparameter combinations across the five guidance signals[1]. To reduce the search space, we conduct our experiments in two phases:

1. We first train models for each pair $(\sigma, \theta) \in \{0, 1, 5, 9, 13\} \times \{0\%, 10\%, 30\%, 50\%\}$ while keeping the interaction probability $p = 100\%$ and using the Concatenation input adaptor. This constrains the parameter space to 20 (σ, θ) pairs, resulting in 55 trained models across the five guidance signals[1].

2. We then fix the optimal (σ, θ) for each guidance signal and train models for all combinations of input adaptors ∈ {Concat, DMF, Conv1S} and interaction probabilities $p \in \{50\%, 75\%, 100\%\}$, yielding an additional 45 models across all guidance signals.

[1] θ is only examined for EDT and GDT, leading to this number of combinations of hyperparameters.

(H1)	σ	$\{0, 1, 5, 9, 13\}$	**(H2)**	θ	$\{0\%, 10\%, 30\%, 50\%\}$
(H3)	Input Adaptor	{Concat, DMF, Conv1S}	**(H4)**	p	$\{50\%, 75\%, 100\%\}$

Table 1: Variation of guidance signal-specific hyperparameters **(H1) – (H4)** in our experiments.

This two-phase approach reduces the brute-force requirement of training 495 models to 100 models, while still systematically exploring the influence of each hyperparameter on performance.

3.2.1.3 *Evaluation Metrics*

In both evaluation phases, we utilize the Dice Similarity Coefficient (DSC), also known as Dice score, to evaluate the segmentation models, defined over a predicted mask X and label Y as:

$$\mathrm{DSC}(X, Y) = 2 \times \frac{|X \cap Y|}{|X| + |Y|} \tag{10}$$

To determine the optimal hyperparameter combination of **(H1) – (H4)** for each guidance signal, we use the Dice score of model predictions after all $N = 10$ user clicks, representing the final segmentation result after annotator interaction. Once the optimal hyperparameters for each guidance signal are identified, we assess their final performance using five complementary metrics, **(M1) –(M5)**, summarized in Table 2. Using a diverse set of evaluation metrics enables a more comprehensive analysis of the guidance signals, capturing their distinct behaviors and accounting for potential trade-offs between factors such as segmentation quality and interaction efficiency.

3.2.2 *Results: Phase 1*

In this phase, we train a *DeepEdit* [67] model for each (σ, θ) pair and set $p = 100\%$ and the input adaptor to Concatenation to constrain the parameter space.

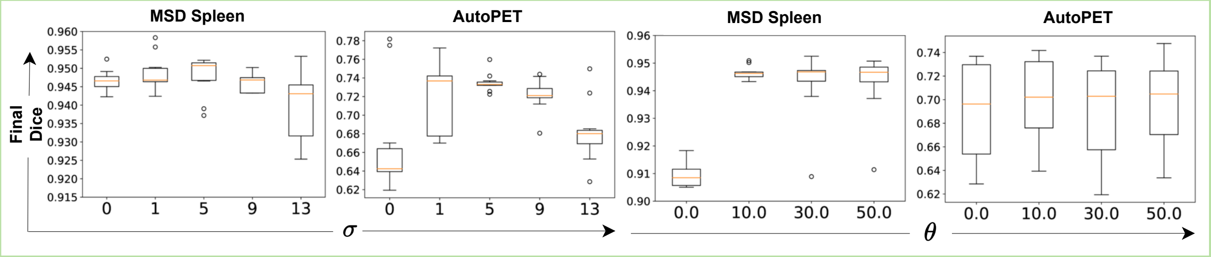


Figure 8: Final Dice scores across different σ (left) and θ (right) aggregated for all five guidance signals.

Metric		Description
(M1)	Final Dice	Mean Dice score after $N = 10$ user clicks per image.
(M2)	Initial Dice	Mean Dice score before any user clicks $(N = 0)$. A higher initial Dice indicates less manual correction required from the annotator.
(M3)	Efficiency	Inverted* time measurement $(1 - T)$ in seconds required to compute the guidance signal. Lower efficiency increases annotation time with each additional click. This metric depends on both volume size and hardware configuration. *The maximum measured runtime T_{max} is shorter than 1 second.
(M4)	Consistent Improvement	Proportion of clicks $\mathcal{C}^+$ that improve the Dice score relative to the total number of validation clicks: $\frac{\lvert\mathcal{C}^+\rvert}{N \cdot \lvert\mathcal{D}_{val}\rvert}$, where $N = 10$ and $\mathcal{D}_{val}$ denotes the validation dataset.
(M5)	Ground-truth Overlap	Overlap between the guidance signal G and the ground-truth mask M: $\frac{\lvert M \cap G\rvert}{\lvert G\rvert}$. This metric estimates the precision of the guidance signal. Since corrective clicks often occur near object boundaries, excessively large guidances (e.g., disks with high σ) may overlap significantly with background regions.

Table 2: Evaluation metrics used to compare guidance signals in our experiments.

(H1) Sigma. As shown in Figure 8, on *MSD Spleen* [163], the highest Dice scores are achieved at $\sigma = 5$. A slight improvement is observed for two samples at $\sigma = 1$, but performance declines for larger values ($\sigma > 5$). On *autoPET* [279], the best performance is seen at $\sigma = 5$ and, for two samples, at $\sigma = 0$, whereas higher σ values again result in a notable drop. Figure 9 shows that for disks and heatmaps, the best initial and final Dice scores are obtained with $\sigma = 1$ and $\sigma = 0$, while geodesic maps perform worse for small $\sigma < 5$ and reach their peak at $\sigma = 5$ on both datasets (green lines). Across all guidance signals, larger σ values consistently reduce initial Dice. The differences between σ values are more pronounced on *autoPET* [279], reflecting the dataset's higher difficulty [82, 239].

Key takeaway: Moderate sigma values $\sigma \leqslant 5$ generally provide the best balance across datasets, while higher values can degrade performance for all signals, especially on more challenging datasets like *autoPET*.

(H2) Theta. We study the effect of truncating large distance values in the EDT and GDT guidance signals (Equations 7 and 8). Figure 8 indicates that the highest final Dice scores on *MSD Spleen* [163] occur at $\theta = 10\%$. On *autoPET* [279], Dice scores are relatively stable across different θ values, with a small improvement at $\theta = 10\%$. Figure 10 further confirms that $\theta = 10\%$ is optimal for both datasets (orange line), and that not truncating distance values on *MSD Spleen* ($\theta = 0\%$) leads to a sharp performance drop.

Key takeaway: A small truncation threshold ($\theta = 10\%$) improves performance, preventing excessively large distances from reducing the effectiveness of guidance signals, particularly on *MSD Spleen*.

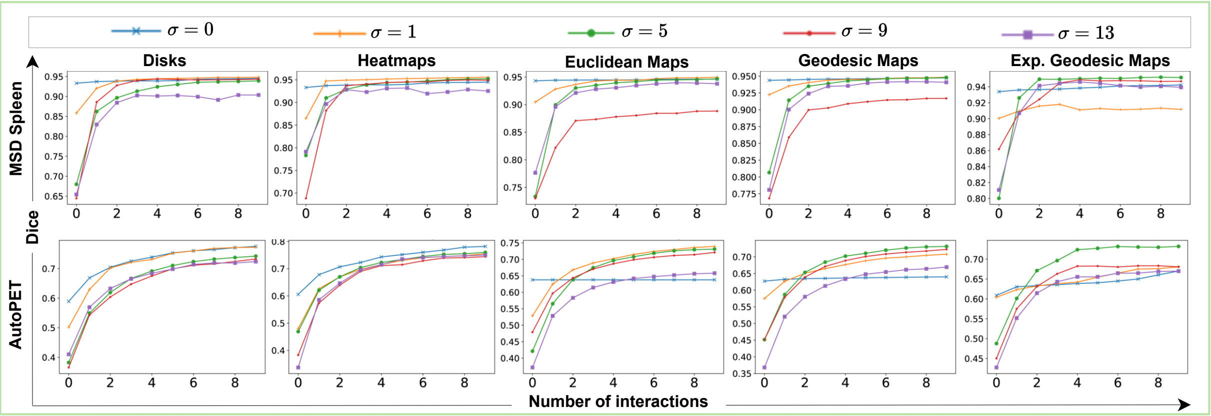


Figure 9: Influence of the σ parameter **(H1)** on the Dice curves for all guidance signals.

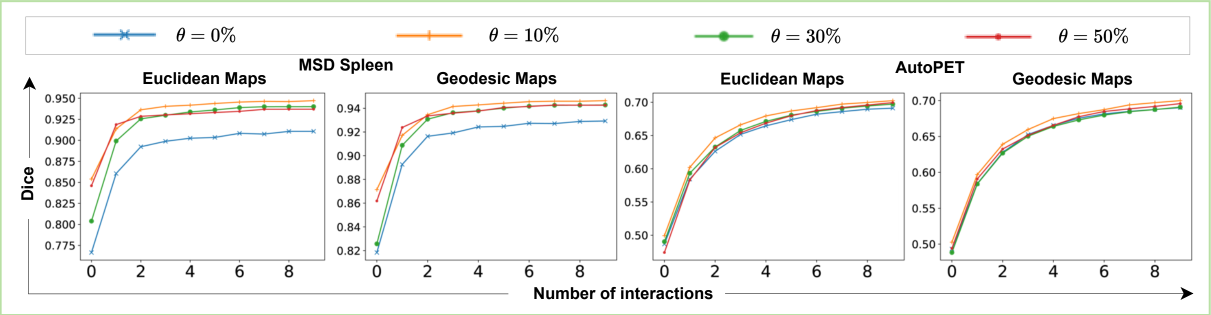


Figure 10: Influence of the θ parameter **(H2)** on the Dice curves for the EDT and GDT guidance signals.

3.2.3 *Results: Phase 2*

In the second phase of our experiments, we fix the optimal (σ, θ) pair for each of the five guidances (see Table 3) and train a *DeepEdit* [67] model for all combinations of input adaptors and probability of interaction.

(H3) Input Adaptor. We evaluate different strategies for combining guidance signals with input volumes using the input adaptors proposed by Sofiiuk et al. [221]. As shown in Figure 11, the best performance across both datasets is achieved by simply concatenating the guidance signal with the input volume. The performance gap between concatenation and the more complex adaptors is substantial, highlighting that simple fusion can be highly effective. A possible reason for this gap is that these adapters are initially proposed for 2D natural images and might not generalize well to the 3D medical image domain [221].

Key takeaway: Concatenation provides a straightforward and robust method for incorporating guidance signals, outperforming more complex adaptors.

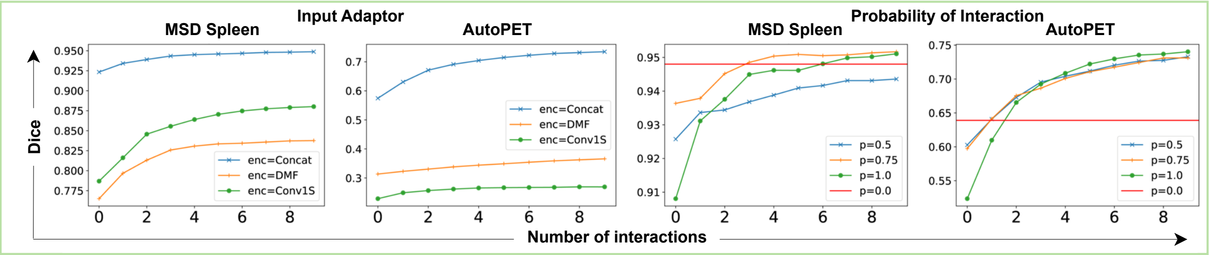


Figure 11: Influence of the **(H3)** input adaptor (left) and **(H4)** probability of interaction (right) hyperparameters on the Dice curves for all guidance signals.

	MSD Spleen [163]							autoPET [154]						
	non-int.	Disks	Heatmaps	EDT	GDT	exp-GDT	Ours	non-int.	Disks	Heatmaps	EDT	GDT	exp-GDT	Ours
σ	-	1	1	1	5	5	adaptive	-	0	0	1	5	5	adaptive
θ	-	-	-	10%	10%	-	-	-	-	-	10%	10%	-	-
Adaptor	Concat	Concat						Concat	Concat					
p	0%	75%						0%	100%					
Dice	94.90	95.91	95.82	95.81	95.19	95.22	**96.87**	64.89	78.15	78.24	75.22	74.50	73.19	**79.89**

Table 3: Optimal parameter settings of our interactive models and the non-interactive baseline (non-int.) and their **(M1)** Final Dice scores.

(H4) Probability of Interaction. Figure 11 illustrates the effect of varying the interaction probability p. On *MSD Spleen* [163], $p \in \{75\%, 100\%\}$ achieves the best final Dice scores, with faster convergence observed for $p = 75\%$. In contrast, $p = 50\%$ performs worse than the non-interactive baseline ($p = 0\%$). On *autoPET* [279], performance is similar across all p values, although the highest Dice is reached at $p = 100\%$. However, $p = 100\%$ leads to lower initial Dice and requires more interactions to converge, indicating that the model relies heavily on interactive inputs.

Key takeaway: A moderate to high interaction probability ($p \in \{75\%, 100\%\}$) generally maximizes performance, but higher p increases dependency on interactions and slows initial convergence. For all subsequent experiments, we use the optimal hyperparameters for each guidance signal summarized in Table 3.

3.2.4 *Addressing Gaps from Phases 1 and 2: Adaptive Gaussian Heatmaps*

The optimal parameters identified in Phases 1 and 2, summarized in Table 3, indicate that disks and heatmaps generally outperform distance-based signals, particularly on the more challenging *autoPET* dataset, where the difference is pronounced. A likely reason is that distance-based signals are rarely used iteratively in prior work [209], as each new click completely modifies the signal, whereas disks and heatmaps simply add a localized, spherical object to the guidance signal. This behavior is illustrated on the bottom row of Figure 12.

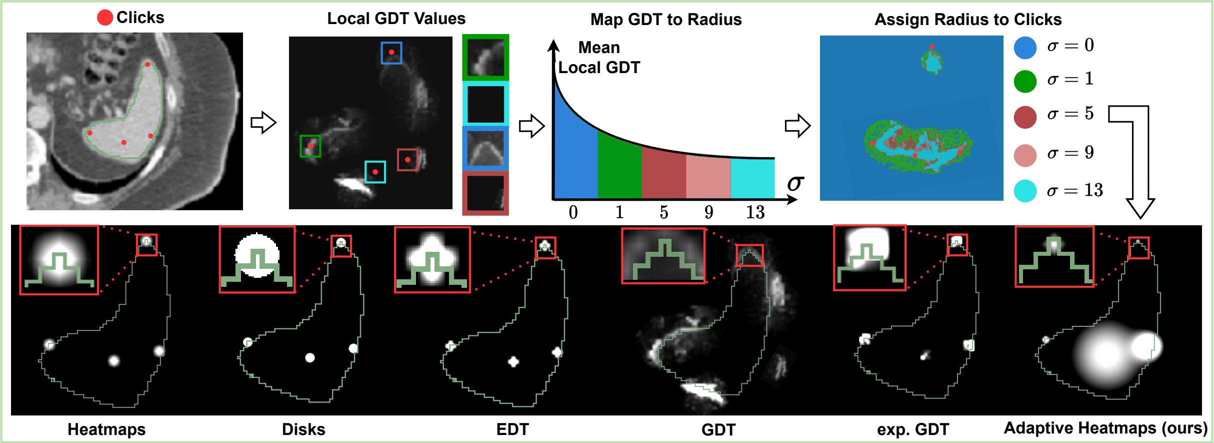


Figure 12: (Top) Example of how our adaptive Gaussian heatmaps determine radii based on different click locations with different local GDT values around each click. (Bottom) Examples of all guidance signals for the same clicks on the *MSD Spleen* dataset [163].

Geodesic-based signals, while potentially informative, are inherently noisy because they depend on the underlying image features (see Figure 12). However, they encode valuable information about object boundaries, which models can leverage, and prior work has demonstrated their effectiveness for non-iterative models [5, 38, 65, 85].

Disks and heatmaps are also limited in *context*, as each new click is encoded using the same spherical structure regardless of the underlying image content. This motivates the design of a hybrid signal that mitigates both the lack of contextual information in heatmaps and the noisy nature of geodesic maps by using the geodesic map to dynamically define the size of each heatmap. While some prior works [73, 223, 278] propose using a larger radius for the first click, our adaptive heatmaps provide greater flexibility by adjusting the radius at each new click based on the underlying image features.

We define our **adaptive Gaussian heatmaps** ad-heatmap(v, c_i, σ_i) via:

$$\sigma_i = \lfloor a e^{-bx} \rfloor, \text{where} \ \ x = \frac{1}{|\mathcal{N}_{c_i}|} \sum_{v \in \mathcal{N}_{c_i}} \mathrm{GDT}(v, \mathcal{C}) \qquad (11)$$

Here, $\mathcal{N}_{c_i}$ is the 9-neighborhood of c_i, $a = 13$ limits the maximum radius to 13, and $b = 0.15$ is set empirically as described in Figure 13. The radius σ_i is smaller for higher x, i.e., when the mean geodesic distance in the neighboring voxels is high, indicating large intensity changes such as edges. This leads to a precise guidance with a smaller radius σ_i near edges and a larger radius in homogeneous areas such as clicks in the center of the object of interest. An example of this process can be seen in Figure 12 on the top row.

We empirically determined the parameters a and b in Equation 11 by analyzing the statistics of local geodesic distances within a $9 \times 9 \times 9$ voxel neighborhood around each click, using the optimal GDT model on the validation set. These statistics are illustrated in Figure 13. The blue bars represent the local GDT values as a function of the click index. Interestingly, later clicks ($N > 6$) tend to correspond to higher GDT values. This behavior arises because, after several interactions, the segmentation mask becomes mostly accurate, and subsequent clicks are typically placed near object boundaries for refinement. In contrast, the initial clicks (1–6) are usually positioned closer to object centers, resulting in lower local GDT values.

To capture this trend, we fitted an exponential curve to the observed local GDT values and set $a = 13$ and $b = 0.15$, reflecting the empirically derived relationship in our guidance signal. Figure 14 illustrates how the fitted curve is used to determine the radius of each new click.

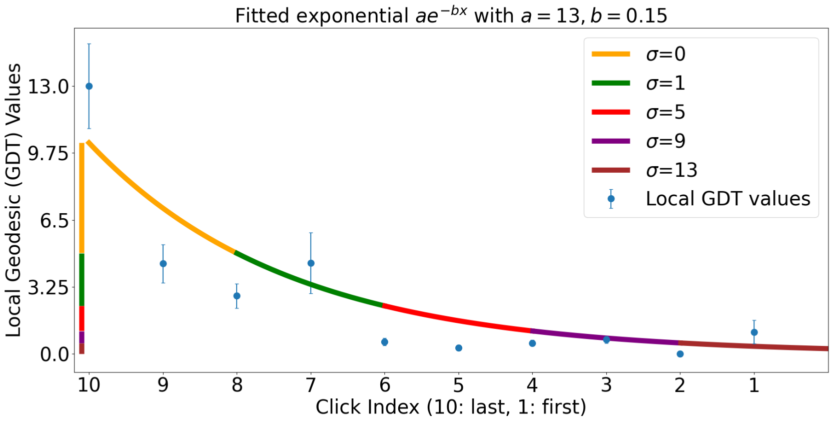


Figure 13: We empirically determine a and b from Equation 11 using our optimal GDT model on the validation split and storing its local GDT values. The x-axis shows click indices, while blue bars show the mean and standard deviation of local GDT values. The colored curve represents an exponential fit to all local GDT values, and the vertical line maps local GDT values to adaptive radii for new clicks in our adaptive heatmaps. A "Local GDT" value refers to the mean GDT value in a $9 \times 9 \times 9$ voxel context around each click.

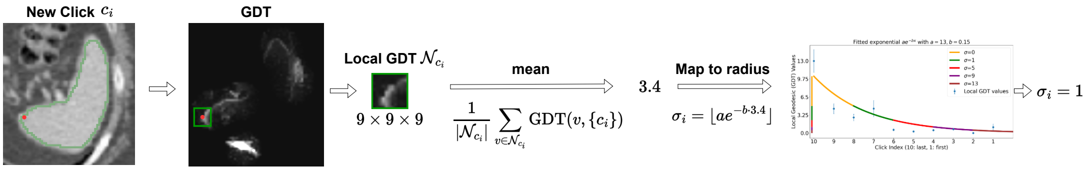

New Click c_i
GDT
Local GDT $\mathcal{N}_{c_i}$
$9 \times 9 \times 9$
mean
$\frac{1}{|\mathcal{N}_{c_i}|} \sum_{v \in \mathcal{N}_{c_i}} \mathrm{GDT}(v, \{c_i\})$
3.4
Map to radius
$\sigma_i = \lfloor a e^{-b \cdot 3.4} \rfloor$
Fitted exponential ae^{-bx} with $a = 13, b = 0.15$
$\sigma_i = 1$

Figure 14: Visual example of Equation 11. A click c_i is added, and the mean of the local GDT values $\mathcal{N}_{c_i}$ around c_i is computed. Finally, the fitted exponential from our best GDT model is used to map the mean local GDT value $x = 3.4$ to a radius $\sigma_i = 1$.

3.2.5 *Comparison of Guidance Signals*

The comparison of the optimal versions of all guidance signals using our five metrics **(M1)–(M5)** can be seen in Figure 15 and Table 4. Although the concrete values for *MSD Spleen* [163] and *autoPET* [279] are different, the five metrics follow the same trend on both datasets.

Guidance Signal	MSD Spleen					autoPET				
	(M1)	**(M2)**	**(M3)**	**(M4)**	**(M5)**	**(M1)**	**(M2)**	**(M3)**	**(M4)**	**(M5)**
Disks	0.95	0.85	0.73	0.52	**0.64**	0.78	0.56	0.62	0.95	**0.31**
Heatmaps	0.95	0.86	0.72	0.42	0.61	0.78	0.60	0.60	0.89	0.25
EDT	0.95	0.89	0.73	0.65	0.46	0.75	0.47	0.68	0.81	0.20
GDT	0.95	0.83	0.66	0.37	0.43	0.73	0.46	0.60	0.22	0.17
exp-GDT	0.95	0.80	0.63	0.39	0.42	0.73	0.43	0.59	0.21	0.04
(Ours) Adaptive Heatmaps	**0.96**	**0.90**	**0.81**	**0.80**	0.20	**0.79**	**0.61**	**0.78**	**0.96**	0.01

Table 4: Exact values for each guidance signal from Fig. 2 from the main manuscript. Best values are in **bold**, second best are underlined. **(M1)**: Final Dice, **(M2)**: Initial Dice, **(M3)**: Consistent Improvement, **(M4)**: Ground-truth overlap, **(M5)**: Efficiency.

(M1) Final and (M2) Initial Dice. Overall, all guidance signals improve their initial-to-final Dice scores after N clicks, with *autoPET* [279] showing a large gap between disks/heatmaps and distance-based signals. Moreover, geodesic-based signals have lower initial scores on both datasets and require more interactions. Our adaptive heatmaps achieve the best scores on both metrics.

(M3) Consistent Improvement. Heatmaps, disks, and EDT achieve the highest consistent improvement on both datasets, which means they are more precise in correcting errors. In contrast, geodesic distances change globally with new clicks as the whole guidance must be recomputed. These changes may confuse the model and lead to inconsistent signals across click iterations. However, our adaptive heatmaps outperform the rest of the guidance signals with a large margin, showing that precise corrections near edges are crucial.

(M4) Overlap with Ground Truth. Heatmaps, disks, and EDT have a significantly higher overlap with the ground truth compared to geodesic-based signals, particularly on *autoPET* [279]. GDT incorporates the changes in voxel intensity, which is not a strong signal for lesions with weak boundaries in *autoPET* [279], resulting in a smaller overlap with the ground truth. The guidances are ranked in the same order in **(M3)** and in **(M4)** for both datasets. Thus, a good overlap with the ground truth may be associated with precise corrections. Additionally, our adaptive heatmaps are able to consistently outperform all other signals on this metric.

(M5) Efficiency. Efficiency is much higher on *MSD Spleen* [163] compared to *autoPET* [279], as *autoPET* has a ×2.4 larger mean volume size. Disks are the most efficient signal, filling up spheres with constant values, while heatmaps are slightly slower due to applying a Gaussian

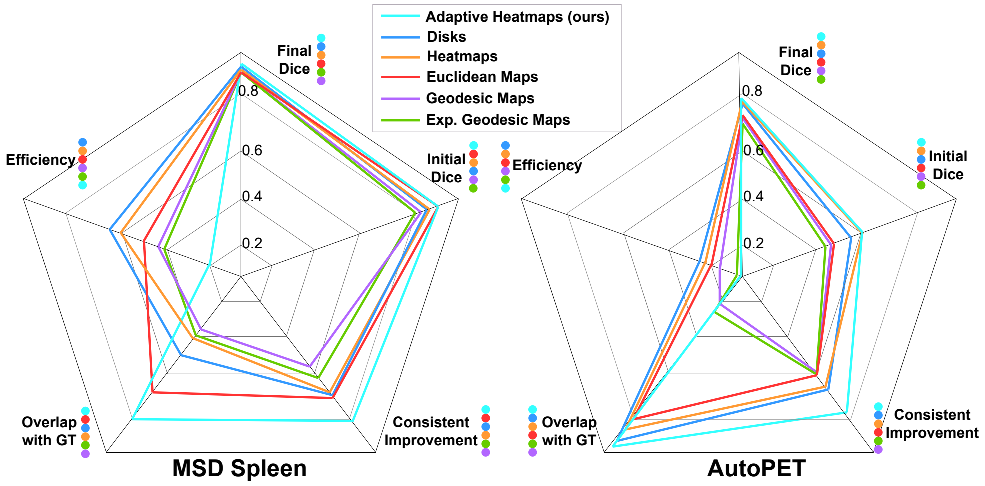


Figure 15: Comparison of all guidance signals with our 5 metrics. The circles next to each metric represent the ranking of the guidances (sorted top-to-bottom).

filter over the disks. Distance transform-based guidances are the slowest on both datasets due to their complexity, but all guidance signals are computed in a reasonable time (< 1s). Our adaptive heatmaps have the worst efficiency as they are a combination of GDT and heatmaps and require the computation of both signals.

Summary of results: Varying **(H1)-(H4)** and examining **(M1)-(M5)**, we find disks and heatmaps as the best existing signals for both tasks, but with inflexibility due to their fixed radius. Using GDT as a proxy signal to adapt the radius of each click mitigates this weakness by imposing large radii in homogeneous areas and small, precise radii near edges. This results in substantially higher consistent improvement and overlap with ground truth and the best initial and final Dice (Table 4). Thus, our comparative study has led to the creation of a more consistent and flexible signal with an improved segmentation, albeit with an efficiency cost due to the need to compute both GDT and heatmaps.

Limitations: Our adaptive heatmaps improve performance across 4 out of 5 metrics but involve extensive computation of both GDT and heatmaps, resulting in low computational efficiency. While we demonstrate that geodesic distances can be applied indirectly for iterative refinement, further optimization is necessary for clinical applicability, where efficiency and rapid responsiveness to user interactions are critical. Another limitation is that our study focuses solely on click-based guidance signals, warranting further exploration of alternative interaction types in future work. Nevertheless, our results highlight the importance of carefully designing and selecting the guidance signal, a factor we show in Chapter 7 to be one of the keys to success for the winning solutions in our co-organized challenges.

3.3 CHAPTER CONCLUSION

This chapter addressed the missing theoretical and practical foundations for representing user interactions in interactive segmentation. We formally defined the concept of a guidance signal (Definition 1), distinguished between explicit and implicit forms, and provided detailed formulations for explicit guidance signals identified in prior work [209] (Equations 4–9). We then systematically investigated how to select an appropriate click-based guidance signal by comparing the most prevalent signal types across two medical segmentation tasks.

To ensure a fair comparison, we optimized each signal's hyperparameters by training 100 interactive models, revealing best practices—most notably, the benefits of conservative values for the radius σ and threshold θ. Using these settings, we evaluated all signals across five complementary metrics capturing segmentation quality, efficiency, and responsiveness to user interactions. Disk- and Gaussian-based signals generally outperformed distance-based alternatives but lacked adaptability to local image context, whereas geodesic-based signals effectively captured boundary information but were noisier and less robust iteratively.

To address these limitations, we introduced adaptive Gaussian heatmaps, a hybrid signal that adjusts each click's radius based on local geodesic distances. This approach combines Gaussian precision with geodesic contextual sensitivity, achieving superior performance across multiple metrics and datasets, particularly in challenging scenarios such as *autoPET* [279]. While enhancing segmentation quality and consistency, this method incurs additional computational cost, highlighting the trade-off between accuracy and efficiency.

Overall, our work establishes a principled framework for the selection, tuning, and design of guidance signals. It offers both theoretical insight and practical recommendations for advancing future interactive segmentation systems.

We summarize the scientific impact of this chapter in four contributions:

Contribution 1: A first formal definition of a guidance signal and its subtypes (implicit and explicit), providing a theoretical foundation for interactive segmentation (in *TPAMI 2024* [209]).

Contribution 2: A comprehensive comparative analysis of click-based guidance signals, including practical recommendations for hyperparameter tuning and insights into each signal's strengths and weaknesses based on 100 trained models (in *MICCAI 2023* [91]).

Contribution 3: The introduction of adaptive Gaussian heatmaps—a novel, context-aware guidance signal that improves segmentation accuracy and iterative correction, outperforming prior approaches on two medical tasks (in *MICCAI 2023* [91])

Contribution 4: An evaluation framework for comparing guidance signals across five metrics that collectively assess segmentation quality, efficiency, and interactivity (in *MICCAI 2023* [91]).

Our contributions highlight the importance of the choice of the guidance signal when designing interactive segmentation models, thereby encouraging the exploration of alternative interaction types (e.g., scribbles, bounding boxes) and other segmentation tasks. The theoretical foundations enable researchers to categorize prior work on interaction representations and choose a suitable signal based on the application. Our adaptive heatmaps outperform prior signals on four out of our five evaluation metrics, but still lag behind on efficiency, due to the complex nature of geodesic-based signals. Hence, an important question remains:

How can we design efficient interactive models?

In the following chapter, we address this question by focusing on two key aspects: (1) improving inference efficiency through an adaptation of the sliding window inferer for interactive segmentation, and (2) enhancing training efficiency by optimizing the computation of distance-based guidance signals, which are commonly used to simulate user interactions for training.

4 EFFICIENT INTERACTIVE MODELS

The second fundamental component we explore in this thesis is the *efficiency* of interactive segmentation models. Responsiveness to user interactions is essential: the system must react quickly to allow the user to inspect predictions and continue interacting in a human-in-the-loop workflow. In Section 4.1 (based on our *ISBI 2024* publication [239]), we focus on improving inference efficiency by adapting interactive models to use sliding window inference during both training and testing. This architecture-agnostic approach can be applied to any interactive segmentation model and offers a significant advantage: during iterative corrections, the model only processes the patch affected by the annotator's interaction rather than the entire image, providing substantial efficiency gains, particularly for volumetric data. Efficient interactive training is another critical aspect. During training, simulating user interactions based on the model's previous predictions can be computationally expensive, as many existing methods rely on distance transforms to determine plausible interaction locations. As discussed in the previous chapter, computing these distance transforms can create a major efficiency bottleneck. Therefore, in Section 4.2, we present methods for efficiently computing distance transforms, reducing training time, and accelerating the generation of distance-based guidance signals.

4.1 EFFICIENT MODEL INFERENCE

This section is based on our publication in *ISBI 2024* [239], © IEEE.

4.1.1 *Global vs. Local Inference*

Interactive segmentation models typically incorporate the guidance signal globally by concatenating it with the image input [10, 38, 140, 209]. In our experiments with the input adaptor hyperparameter **(H3)** in Chapter 3, we also found this to be the most effective way to integrate click-based explicit signals. However, unlike bounding box or scribble-based inputs, click interactions are inherently *local*, providing precise information about the user's intent. Traditional global inference strategies, where the entire image and guidance signal are provided to the model in full, do not align optimally with the localized nature of click-based interactions.

One solution to this issue is to use local inference techniques, where the model only sees a subset of the input at a time. A widely adopted approach in medical image segmentation is sliding window inference [135, 280, 281], which partitions the image into fixed-size patches and makes individual predictions for each patch. These patch-level predictions are then aggregated through window stitching to produce the final output, with overlapping regions combined using an ensembling mechanism. Sliding window inference is particularly advantageous for large inputs that cannot fit entirely into memory, as processed patches can be offloaded to the CPU while the model continues processing remaining patches. In contrast, global inference requires either resizing the volume via interpolation or cropping, both of which can lead to a loss of detail [282]. Figure 16 illustrates the application of global (top) and local (bottom) inference strategies to large volumes, such as whole-body PET scans, in the context of interactive segmentation.

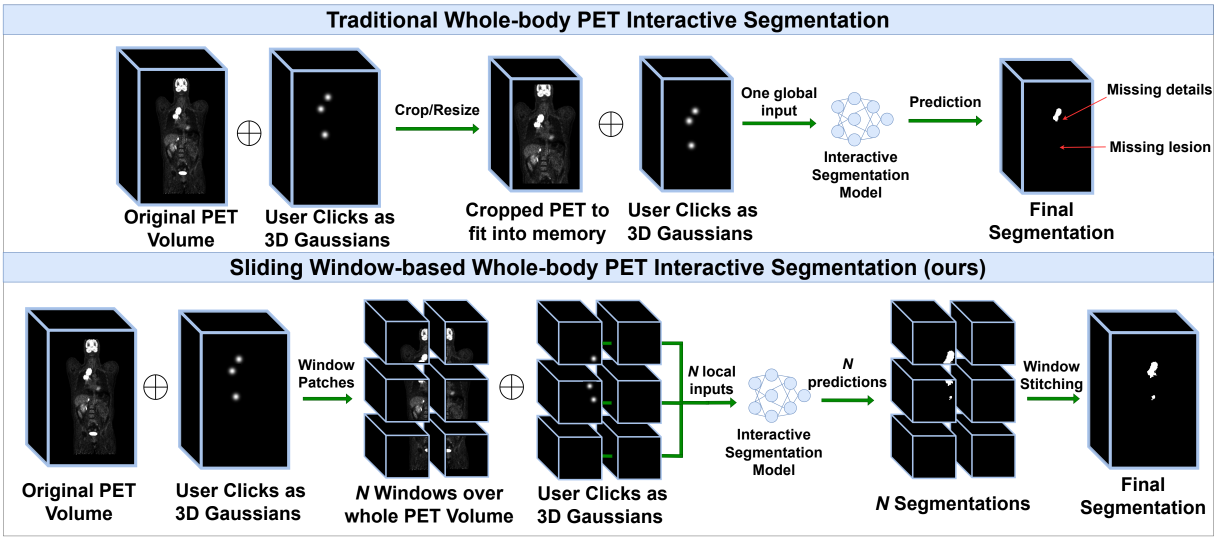

Traditional Whole-body PET Interactive Segmentation
Original PET Volume
User Clicks as 3D Gaussians
Crop/Resize
Cropped PET to fit into memory
User Clicks as 3D Gaussians
One global input
Interactive Segmentation Model
Prediction
Missing details
Missing lesion
Final Segmentation
Sliding Window-based Whole-body PET Interactive Segmentation (ours)
Original PET Volume
User Clicks as 3D Gaussians
Window Patches
N Windows over whole PET Volume
User Clicks as 3D Gaussians
N local inputs
Interactive Segmentation Model
N predictions
N Segmentations
Window Stitching
Final Segmentation

Figure 16: Comparison of global inference (top) and local inference using the sliding window strategy (bottom). Global inference requires cropping or resizing large images, which can cause information loss. In contrast, sliding window inference processes the image patch by patch and aggregates predictions using window stitching, preserving spatial detail.

Nevertheless, sliding window inference introduces significant computational overhead, as the model must generate predictions for multiple patches, often sequentially due to memory constraints, which slows down inference. While this is unavoidable for non-interactive models where all patches are equally relevant, there is a straightforward solution for click-based interactive segmentation. In a human-in-the-loop setting, each new click targets a specific region. Therefore, it is unnecessary to evaluate all patches; the model only needs to update predictions in the patch containing the click. Predictions for other patches can be left unchanged, as they are largely independent[1]. This strategy is illustrated in Figure 17. Note that the first prediction must be done for all patches to obtain the initial segmentation mask. All subsequent corrective predictions can then be performed on only the clicked patch (or patches).

[1] With the exception of clicks in regions where patches overlap. Then, each overlapping patch prediction needs to be updated.

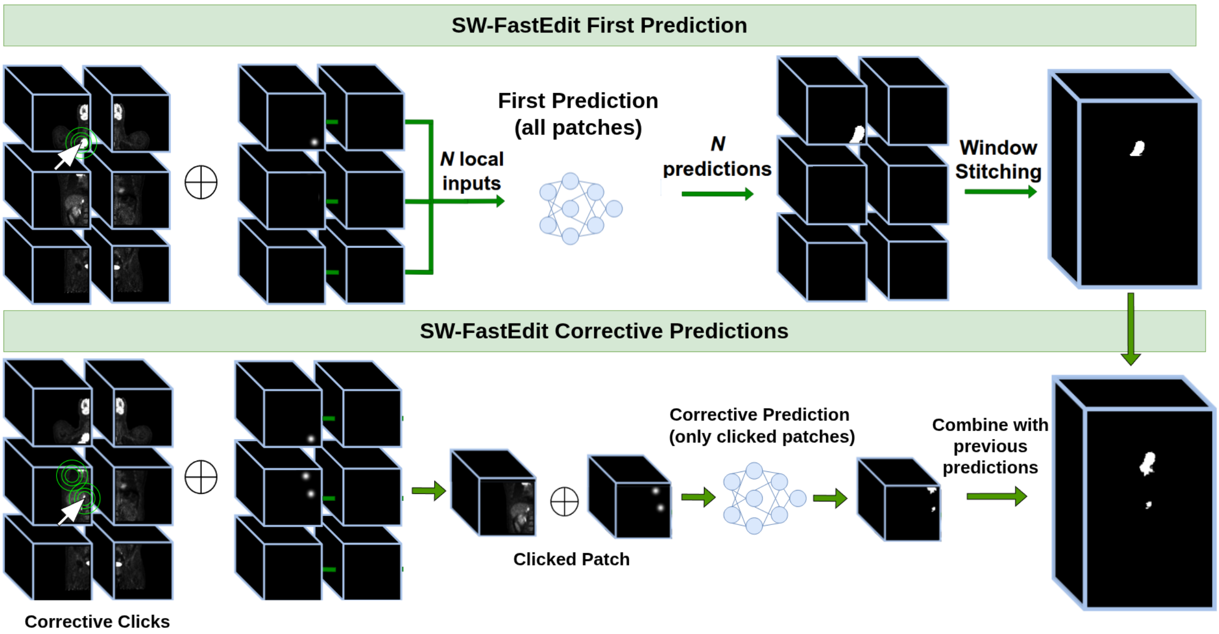

SW-FastEdit First Prediction
N local inputs
First Prediction (all patches)
N predictions
Window Stitching
SW-FastEdit Corrective Predictions
Corrective Prediction (only clicked patches)
Combine with previous predictions
Clicked Patch
Corrective Clicks

Figure 17: Sliding window-based interactive segmentation. In the first prediction (top row), the model needs to predict for all patches to produce an initial mask. However, in corrective predictions (bottom row), the model only needs to address the clicked patches.

This approach provides a substantial efficiency gain, since patches are typically much smaller than the full image, where common patch sizes are (96 × 96 × 96), (128 × 128 × 128), and (192 × 192 × 192) [67, 135, 231, 281], whereas full image volumes are much larger, e.g., the mean image size in *autoPET* [279] is (400 × 400 × 352), and in *MSD Spleen* [163] it is (512 × 512 × 89). Building on this idea, we propose a more efficient interactive model, *SW-FastEdit* (as described in our *ISBI 2024* publication [239]), which extends the *DeepEdit* model [67] using our codebase from Chapter 3 to support sliding window-based training and inference. In the following sections, we describe its interactive training and inference mechanisms and compare its performance with the standard *DeepEdit* model that uses global inference over the whole image.

4.1.2 *Experimental Setup*

Datasets and Pre-processing. We use the *autoPET* dataset [279] for training and validation, containing whole-body PET/CT scans of patients with lung cancer, melanoma, or lymphoma. We only use PET volumes of unhealthy patients and use an 80-20 training-validation split. We normalize the PET volumes and clip values beyond the 0.05th and 99.95th percentiles. During training, we randomly crop the image to a subvolume with a size of (224 × 224 × 224) voxels and a probability $p_{\text{lesion}} = 60\%$ to be centered around a lesion and $p_{\text{bg}} = 40\%$ around a non-lesion voxel. We apply random flips along each axis with a 10% probability and random 90° rotations with a 10% probability per axis. We then feed the transformed input to our sliding window-based model with a patch size of (128 × 128 × 128) and a patch overlap of 25% with Gaussian weighting on overlapping patches for the window stitching.

Click Simulation. We train all our models using simulated click corrections and investigate three click generation strategies as an ablation study. Specifically, we either: (1) simulate all clicks at once and provide them directly to the model without an interaction loop; (2) simulate N lesion and N background clicks sequentially in a loop of N iterations based on the over- and undersegmented errors as in Sakinis et al. [10]; or (3) apply (2) but stop simulating clicks if the Dice score for the volume has already reached $\text{Dice}_{\text{max}} = 90\%$. We encode clicks as 3D heatmaps with a radius of one voxel ($\sigma = 1$) as we found this to be the optimal value in Chapter 3, and concatenate them to the PET images as an additional channel, as depicted in Figure 16. During the first prediction step, this channel is empty, and with each simulated corrective click, a new heatmap is generated within this input channel. Note that we simulate a "lesion" and "background" click in each prediction step to correct under- and oversegmentation, respectively, by sampling from the distance transform of the error regions as in Sakinis et al. [10]. This leads to two additional input channels that are not illustrated in Figure 16 for clarity. We set $N = 10$ as the maximum number of click iterations.

Model Optimization. As sliding window inference is not differentiable by default, we utilize *MONAI*'s [283] differentiable implementation of the SlidingWindowInferer for the first time for interactive segmentation, where predictions on overlapping windows are weighted based on the error from each prediction and then averaged so that gradients are linearly combined. We further utilize a cosine annealing learning rate scheduler with an initial learning rate of 10^{-4} and train our models for 200 epochs. We utilize the same architecture from Chapter 3 based on U-Net [127] with 6 down- and upsampling layers.

4.1.3 *Results with Simulated Interactions*

We first evaluate all models with simulated clicks as described in our experimental setup in Section 4.1.2. Then, in Section 4.1.4, we validate *SW-FastEdit* with real interactions from medical annotators.

4.1.3.1 *Evaluation of Inference Time and Memory Usage*

We first investigate how much efficiency we can gain by incorporating the sliding window inference into the interactive loop. We evaluate both *DeepEdit* [67], which uses the whole image as an input, and our model *SW-FastEdit* that uses sliding window inference. As *SW-FastEdit* requires a full forward pass over all sliding window patches for its first prediction, we evaluate both its first prediction and corrective predictions. For *SW-FastEdit*'s first prediction, we process four patches in parallel on the GPU, then offload them to the CPU before processing the next set of four patches, repeating this cycle until the entire image is covered for a single forward pass.

Figure 18 shows the inference time and GPU memory usage of both models on the *autoPET* dataset, with images sorted in ascending order of volume to highlight trends. The blue points represent *DeepEdit*, whose inference time and memory usage scale approximately linearly with the number of voxels. This linear scaling leads to catastrophic failure at around 70 million voxels (e.g., a volume of size (400 × 400 × 448)), where *DeepEdit* exceeds the 40 GB GPU memory limit, resulting in an Out of Memory (OOM) error. These failures are indicated by the red crosses in Figure 18 and account for approximately 8% of all *autoPET* images. Consequently, *DeepEdit* cannot be applied directly to this dataset without interventions such as cropping or resizing, as illustrated in Figure 16.

In contrast, our sliding window-based model, *SW-FastEdit*, achieves roughly ×3 faster inference on its first prediction compared to *DeepEdit*, and a constant inference time of < 0.05 seconds for subsequent corrective clicks. This results in a brief warm-up period for the first prediction, followed by consistently fast responses for all follow-up corrections. As shown on the right side of Figure 18, *SW-FastEdit*'s GPU memory usage remains constant, independent of the input volume, for both first and corrective predictions. The memory footprint is slightly higher for the first prediction because we compute four patches in parallel to accelerate the initial inference, consuming roughly four times more memory than subsequent predictions. Users can increase the number of parallel patches to further speed up the first prediction, but this will proportionally increase memory usage.

Key Takeaway: By combining sliding window inference with local patch predictions, our approach achieves a significant improvement in efficiency for interactive segmentation models. Its constant memory footprint and rapid inference for corrective interactions ensure stable, responsive performance, making it suitable for inputs of any size.

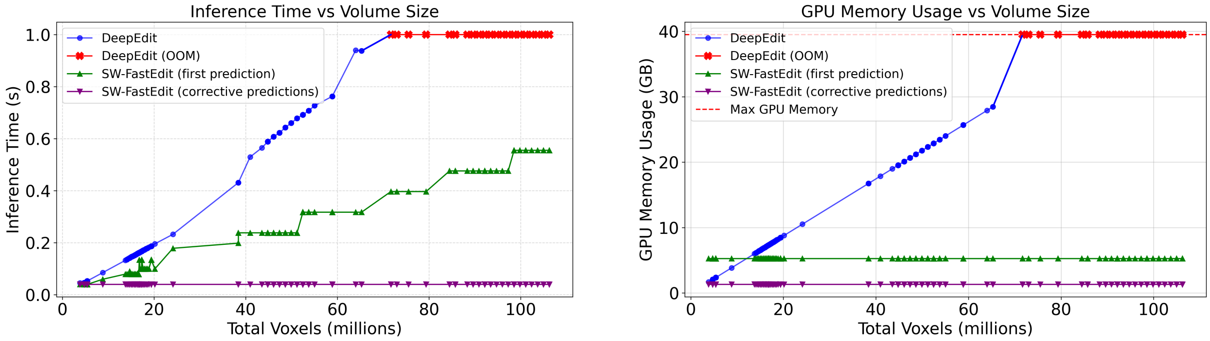


Figure 18: Comparison of inference time (left) and GPU memory usage (right) of *SW-FastEdit* and *DeepEdit* [67] (global inference) across different input sizes from the *autoPET* dataset [279]. Each point corresponds to a volume from the *autoPET* dataset. OOM: Out of memory.

4.1.3.2 *Evaluation of Segmentation Performance*

Although this chapter focuses on evaluating efficiency, we also briefly examine whether switching from global to local inference affects segmentation performance. Since *DeepEdit* cannot fit ≈ 8% of all *autoPET* volumes into memory, we crop all validation volumes to (224 × 224 × 224) using a center crop to enable a fair comparison. Table 5 shows that *SW-FastEdit* maintains performance relative to *DeepEdit*, indeed, it even improves Dice scores by about 1% on the cropped inputs and achieves similar scores on the original full volumes. The difference between cropped and full-volume scores arises because cropping discards portions of the labels.

Key Takeaway: Introducing local inference in *SW-FastEdit* does not come with a sacrifice of the segmentation performance.

	SW-FastEdit (ours)	DeepEdit [67]
Type of Inference	Local (Sliding Window)	Global (Full Volume)
Dice@10 (224^3 crop)	**85.55%**	84.34%
Dice@10 (full volumes)	84.90%	does not fit on 40 GB GPU

Table 5: Comparison between the sliding window and standard inference. Dice@X is the Dice score after X clicks.

4.1.3.3 *Ablation Studies*

We conducted two ablation studies to investigate: (1) how the click simulation strategy during training affects final performance, and (2) whether non-interactive training can improve predictions when no clicks are provided.

Click Generation Strategy (Training)	Validation Dice Score
Non-corrective N = 10 clicks	47.00%
Corrective N = 10 clicks	84.90%
Corrective N = 10 clicks with $\text{Dice}_{\text{max}} = 90\%$	**85.34%**

Table 6: Results (Dice score) for SW-FastEdit when using different click generation strategies during training. During validation, we always simulate exactly 10 clicks for each sample using the simulation strategy of Sakinis et al. [10].

Table 6 summarizes the three click simulation strategies we explored for training *SW-FastEdit*. Simulating all clicks at once, without an interaction loop, leads to a substantial performance drop (47% Dice) compared to the other strategies. Surprisingly, it even performs worse than the non-interactive baseline trained without any clicks (73.04% Dice), likely due to overfitting caused by the two additional input channels. In contrast, training with interaction loops, where the model performs multiple predictions and clicks are simulated iteratively based on its errors, achieves Dice scores of 84.90% and 85.34%. Stopping click generation once the training Dice for an image reaches 90% provides only a marginal improvement.

Key Takeaway: Iterative click simulation during training is crucial for achieving high performance in interactive segmentation models.

Training			
400 epochs without clicks		✓	✓
200 epochs with clicks	✓		✓
Dice@0	24.47%	73.04%	68.03%
Dice@10	85.34%	-	84.79%

Table 7: Results from non-interactive pre-training. Dice@X denotes the Dice score after X clicks.

We also investigated whether non-interactive pre-training improves the model's predictions when no clicks are provided. Table 7 shows that *SW-FastEdit* initially produces poor-quality predictions without clicks, achieving only 24.47% Dice. This is likely due to the model's dependency on interactions, as clicks are always present during training. Motivated by this, we pre-trained *SW-FastEdit* without interactions for 400 epochs, and then continued training for 200 epochs using the standard interactive setup with clicks. While this slightly reduced its final performance after 10 click iterations to 84.79% Dice, it significantly improved zero-click predictions to 68.03% Dice, which is much closer to the non-interactive baseline that achieves 73.04% Dice with 400 epochs.

Key Takeaway: Non-interactive pre-training substantially improves initial segmentation quality when no clicks are provided, i.e., has a higher Dice@0, while maintaining nearly the same performance when subsequent interactive clicks are used.

4.1.4 *Results with Real Interactions from Medical Annotators*

To validate *SW-FastEdit*'s usability in clinical practice, we conducted a user annotation study with one radiologist, one medical doctor, and two medical students in collaboration with the Annotation Lab Essen[2]. The annotators were instructed to annotate the same 10 PET volumes and perform the following loop (1)-(3) exactly 10 times for each volume: (1) predict with the *SW-FastEdit* model; (2) add one lesion click; (3) add one background click. During steps (2) and (3), annotators corrected areas where our model had missegmented in step (1). We used the *SW-FastEdit* model with non-interactive pre-training from Table 7. We assessed the user study using the Dice score, as well as the perceived NASA-TLX workload [147] in Table 8.

[2] https://annotationlab.ikim.nrw/

Annotator	Dice@10 ↑	NASA-TLX [147] ↓
A1 (medical doctor)	**72.49**% ± 18.66%	5.50
A2 (medical student)	64.66% ± 23.13%	3.30
A3 (radiologist)	67.72% ± 21.00%	2.83
A4 (medical student)	65.65% ± 26.24%	2.83
Simulated user@0	51.43% ± 25.21%	-
Simulated user@10	**78.50**% ± 14.96%	-
Non-interactive model	61.69% ± 20.53%	-

Table 8: Results from the user study on 10 unseen volumes from the validation set. The simulated user is the one described in Sakinis et al. [10].

Table 8 presents the user study results. The simulated user achieves a Dice score of 78.50% with 10 clicks. Note that the simulated user always simulates valid clicks since he has access to the ground truth labels. The best annotator (A1) achieves a slightly lower Dice than the simulated user, however, his results are significantly better than the non-interactive model. This shows that an interactive model can deliver much higher quality results than non-interactive models alone. Following the user study, we asked the annotators to fill out a NASA-TLX form, three Likert-scale questions, and one open question for feedback. The three Likert questions were in the scale of 1 to 10: (1) *Are the background clicks necessary to achieve good results?* (2) *Are 10 clicks enough for the annotation?* and (3) *Does SW-FastEdit speed up your annotation time?*

The annotators ranked the mental (3.5/10), physical (2/10), and temporal demand (3.3/10), as well as the effort (4.3/10) and frustration (2.3/10), on average as low. They also rated their performance on the task relatively high (6.5/10). Additionally, the annotators rated 10 click iterations as sufficient (6.6/10), background clicks as necessary (6.5/10), and that *SW-FastEdit* speeds up the annotation (7.5/10), with A3 commenting: "majority of the cases can be annotated with only 3-4 updates, which is really great. The process was much faster than annotating PET images from scratch". Overall, the feedback is positiv,e and the annotators saw potential in applying *SW-FastEdit* in their annotation workflow.

One unusual finding is that the simulated user achieved a much higher performance than all the annotators who used the model, although it also used exactly 10 click iterations, the same *SW-FastEdit* model, and the same validation images. We dive much deeper into the reason for this in Chapter 5 where we investigate how to realistically simulate user interactions to reproduce the performance we actually see when providing such models to medical annotators.

Key Takeaway: *SW-FastEdit* shows that incorporating sliding window inference enhances the performance of previous non-sliding window models, efficiently handling large volumes without requiring resizing or cropping. Our approach demonstrates promising results, with our user study indicating a low perceived NASA-TLX workload with medical annotators expressing favorable opinions and indicating a willingness to use it.

4.2 efficient model training

This section is based on our implementation of the evaluation pipeline for the *CVPR 2025 Foundation Models for Interactive 3D Biomedical Image Segmentation Challenge*[3]. Zdravko Marinov conceptualized and implemented ***all*** components presented in this section.

4.2.1 *Distance Transforms for Interaction Simulation*

In the previous section, we highlighted that iterative training is essential for achieving strong interactive performance, as it effectively teaches the model to respond to user interactions [10, 64, 67, 80, 209]. Many prior approaches simulate user clicks by assuming annotators would click near the object's center [36, 64, 80, 91, 209]. However, objects are not always convex, and the geometric center may lie outside the object. For example, the center of a ring lies in its hole rather than on the ring itself (Figure 19). To address this, several methods incorporate the Euclidean distance transform (EDT) as an additional signal, since higher EDT values[4] correspond to points well within the object mask regardless of its shape.

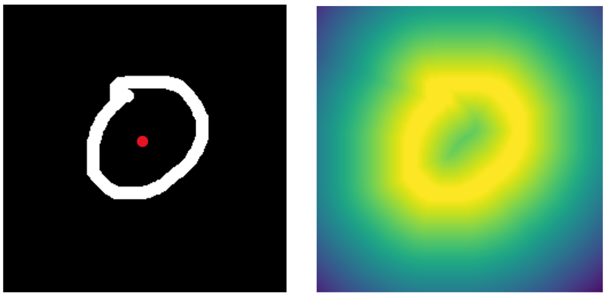


Figure 19: Example of a simulated click in the geometric center of a ring-shaped mask (left) and an EDT computed on the same mask (right). High EDT values are plausible click locations.

One of the most established methods for simulating clicks from the EDT was proposed by Sakinis et al. [10] and has been widely adopted in prior work [23, 36, 49, 52, 64, 67,

[3] https://www.codabench.org/competitions/5263/

[4] In this section, we use the inverted EDT to replicate the implementation of Sakinis et al. [10].

68, 80, 83, 91]. The method involves two steps: (1) computing the inverted EDT[5] for over- and undersegmented errors, and (2) sampling a click using the normalized EDT values as pseudo-probabilities. We introduce simple optimization techniques for both steps to reduce the training time of interactive models.

4.2.2 *Localized Computation of Distance Transforms*

We identify three main opportunities for optimization in the method of Sakinis et al. [10]. First, since we aim to select clicks within the central components of the mask, we can choose one of the local minima instead of sampling from the EDT as pseudo-probabilities, which is inherently slow. Second, distances outside the binary mask are irrelevant for click placement, so their computation can be skipped. Finally, our third optimization is to compute the EDT on a downsampled version of the mask and then map the coordinates back to the original resolution. This is particularly beneficial when processing large volumes that may not fit into GPU memory. Visual examples of these optimizations are shown in Figure 20.

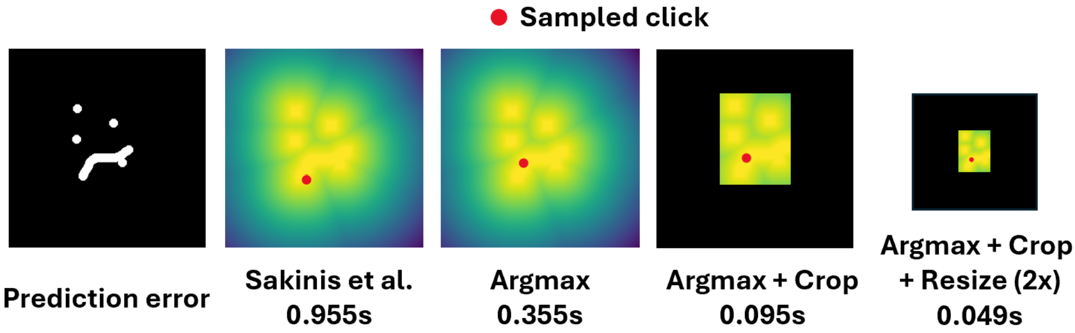


Figure 20: Examples of EDT-based click simulation methods. Given a prediction error mask, the Euclidean Distance Transform (EDT) is computed, and clicks are sampled based on the distances in the EDT. The various strategies show significant differences in computational time.

To assess the efficiency improvements, we computed the EDT and simulated a single click for 100 labels from the *autoPET* dataset [279] using four different click simulation strategies. Figure 21 visualizes the computation times for these cases. The results show that replacing pseudo-probability sampling with an argmax reduces computation time by a factor of three. Further, computing the EDT only over the cropped binary mask provides an additional threefold speed-up. Finally, downsampling the volume by a factor of 2 along each dimension decreases the average click generation time from 1 second to 0.05 seconds, achieving an overall 20× speed-up over Sakinis et al. [10]. Table 9 summarizes the exact numbers. We use CuPy's [284] GPU implementation for computing the EDT and use an A100 GPU to conduct this experiment.

[5] In practice, Sakinis et al. compute the Chamfer distance transform (CDT) as an approximation of the EDT. CDT can be viewed as a discrete version of the continuous EDT.

EDT-based click simulation	Time to simulate a single click in seconds
Sakinis et al. [10]	0.9550 ± 0.4522s
Argmax	0.3555 ± 0.1044s
Argmax + Crop	0.0959 ± 0.0227s
Argmax + Crop + Resize	0.0492 ± 0.0144s

Table 9: Comparison between the EDT-based click simulations on the 100 *autoPET* [279] volumes.

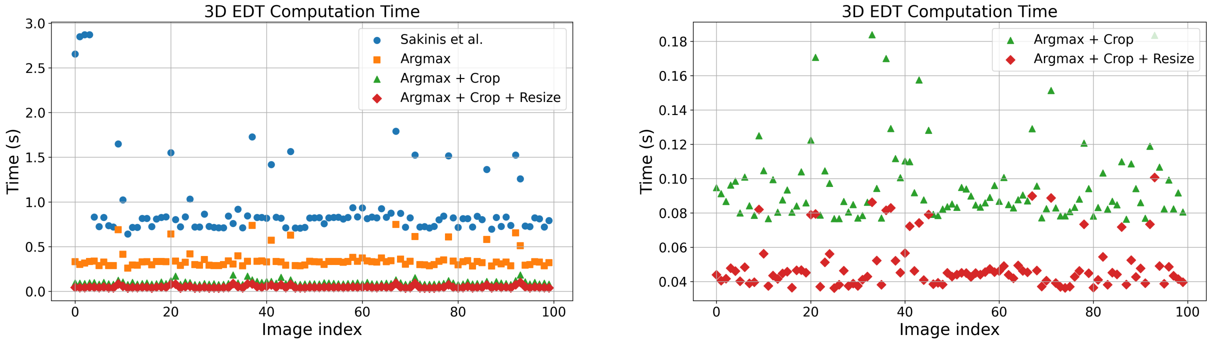


Figure 21: Comparison of techniques for simulating a click from the EDT on 100 volumes from the *autoPET* dataset [279]. The right plot is a zoomed-in version of the left one.

We note that these optimization strategies do not alter the final EDT, as the distances within the cropped region are identical to those computed over the full image. The only minor differences may arise from interpolation artifacts when downsampling the binary mask for the final optimization step.

Key Takeaway. What does this mean in practice? Suppose we train an interactive model using 10 simulated clicks per image. With 500 training images over 200 epochs, this results in 500 × 200 × 10 EDT-based click simulations, equivalent to roughly 11.5 days of training on an A100 GPU using the method of Sakinis et al. [10]. Simply replacing sampling with an argmax reduces this time to 4 days, and applying all of our optimizations further cuts it down to approximately 14 hours - a 20× speed-up in training. This dramatic reduction not only enables training more models and performing extensive ablation studies but also significantly lowers the carbon footprint of model development.

4.3 CHAPTER CONCLUSION

This chapter focused on fundamental improvements in the training and inference efficiency of interactive models, which are major bottlenecks when deploying such models in real-world applications, where fast responses to user interactions are crucial.

We proposed *SW-FastEdit*, a novel approach integrating sliding window inference for interactive segmentation. By leveraging knowledge of which window patches are clicked by the user, we can save computational time during corrective clicks. This simple strategy allows our model to handle inputs of any size while maintaining practically constant memory usage and inference time, in contrast to traditional global models, which often encounter out-of-memory errors for large volumes and slow down as input size increases. We validated *SW-FastEdit* in a real user study with medical annotators, who reported positive feedback and expressed willingness to adopt such models in their daily workflow.

We also explored ways to reduce the training time of interactive models by optimizing the computation of distance transform-based interaction simulations. By focusing on local computation, we achieved a 20× speed-up, reducing the training time of interactive models from, for example, two weeks to just half a day. This not only lowers the carbon footprint but also enables more extensive experimental exploration.

Our contributions to inference and training efficiency are general and architecture-agnostic, making them a fundamental improvement to the entire interactive segmentation paradigm.

We summarize the scientific impact of this chapter in three key contributions:

Contribution 1: *SW-FastEdit* – a sliding window-based interactive segmentation model that leverages local inference to ensure constant inference speed and memory usage, enabling fast responses to user input (in *ISBI 2024* [239]).

Contribution 2: A general strategy to reduce the training time of interactive models via localized computation of distance transforms, achieving a 20× speedup over traditional interactive training paradigms.

Contribution 3: Fundamental improvements to the interactive segmentation paradigm. The local inference strategy in *SW-FastEdit* and our training optimizations are broadly applicable across the majority of interactive model architectures

Our contributions push interactive segmentation models toward more efficient practices, enhancing their responsiveness to user interactions and significantly reducing required training time. Importantly, our results are not tied to any specific model architecture, enabling progress across the entire field. During our user study with *SW-FastEdit*, we observed that the model performed substantially better with simulated interactions than with any of the human annotators. This surprising finding led us to ask:

How can we realistically simulate human interactions?

In the next chapter, we tackle this question by investigating the "human" aspect of data annotation, focusing on inter-annotator disagreement — a key factor behind the discrepancy between real and simulated interactions. We analyze where annotators disagree and use these insights to inform our simulation methods, better reflecting clinical reality and achieving higher realism in the final results.

5

INTERACTION SIMULATION

In this chapter, we investigate how to realistically simulate user interactions to better approximate how interactive models are used in real clinical practice. In Section 5.1 (based on our *TPAMI 2024* publication [209]), we formally define the concept of a *robot user*, introduce its two main types (iterative and non-iterative), and describe their roles in the training and evaluation of interactive models. Section 5.2, based on our *MICCAIw 2024* publication [285], examines how traditional interaction simulation methods compare with real annotator behavior in whole-body PET lesion segmentation. Our analysis shows that conventional robot users consistently yield overly optimistic results, outperforming those obtained with real annotator interactions. Motivated by these surprising findings, we propose a novel robot user that generates more realistic interactions by modeling inter-annotator disagreement and incorporating human factors such as click variability. Through two independent user studies, each involving four annotators, we demonstrate that our approach significantly reduces the discrepancy between simulated and real interactions compared to traditional robot users.

5.1 WHAT IS A ROBOT USER?

This section is based on parts of our publication in *TPAMI 2024* [209], © IEEE.

5.1.1 *Definition of a Robot User*

We aim to unify the terminology used to describe interaction simulation, user prompt generation, or synthetic user modeling under a single term — the robot user. This somewhat archaic term was originally coined 15 years[1] ago [122, 286], when interactive models were primarily based on classical methods such as shape priors [286] or Gaussian Mixture Models [141]. Here, we pay homage to the pioneers of interactive segmentation and reintroduce the term to establish a unified and modern definition in our *TPAMI 2024* publication [209].

[1] Carsten Rother is a pioneer of the interactive segmentation field with *GrabCut* [141] and introduced the term robot user in 2010 [122, 286]. One of his later PhD students, Alexander Kirillov, went on to develop the Segment Anything Model (SAM) [137] at Meta, continuing this line of groundbreaking research in interactive segmentation.

Definition 2: A *robot user* [122] is a simulated model designed to mimic the behavior of a real human annotator. It uses ground-truth labels to generate user interactions at plausible locations. For instance, interactions can be sampled from the ground-truth labels or placed at the center of the largest object. These simulated interactions are then converted into a guidance signal, which the model incorporates for its predictions. Robot users are particularly useful for training on large datasets, where obtaining real human interactions for each image would be impractical. They can also be employed during testing to evaluate models without requiring real human interactions.

Robot users can be categorized as non-iterative or iterative. Non-iterative robot users simulate all interactions simultaneously, integrate them into the model, and perform a single prediction. In contrast, iterative users simulate interactions in a loop. In this case, the model predicts, interactions are generated based on the errors of this prediction, and the model predicts again using all the previous interactions in an interaction-prediction loop [215]. An iteration, e.g., in iterative robot users, denotes a single round of interaction and prediction with the model.

5.1.2 *Non-iterative Robot Users*

Non-iterative robot users generate all interactions from the ground-truth annotation in a single step, without iteratively simulating more interactions based on the model's prediction. These methods can be categorized by the type of simulated interaction: clicks, scribbles, bounding boxes, or polygons. They can also be further divided into sampling-based (stochastic) or rule-based (deterministic) methods.

5.1.2.1 *Click-based Non-iterative Robot Users*

Sampling-based. Clicks, denoted as $\mathcal{C}$, can be simulated by randomly sampling points from the ground-truth mask $\mathcal{L} \in \{0,1\}^{H \times W \times K}$ for 3D images or $\mathcal{L} \in \{0,1\}^{H \times W}$ for 2D images. Sampling strategies vary widely and include uniform sampling [2, 25], distance transform-weighted sampling [10, 67], center-weighted sampling [55, 66], among others [209]. These strategies can be formally summarized as:

$$\mathcal{C} = \{c \mid c \sim P(\mathcal{L})\}, c \in \mathbb{N}^D, D \in \{2,3\} \tag{12}$$

where $P(\mathcal{L})$ denotes a probability distribution over the ground-truth pixels, which may be uniform, boundary-biased, center-biased, or follow other weighting schemes [209], and D is the image dimensionality.

Rule-based. Deterministic click strategies place clicks at locations expected to provide meaningful guidance. Common approaches are object centers [3], extreme points [176], and interior

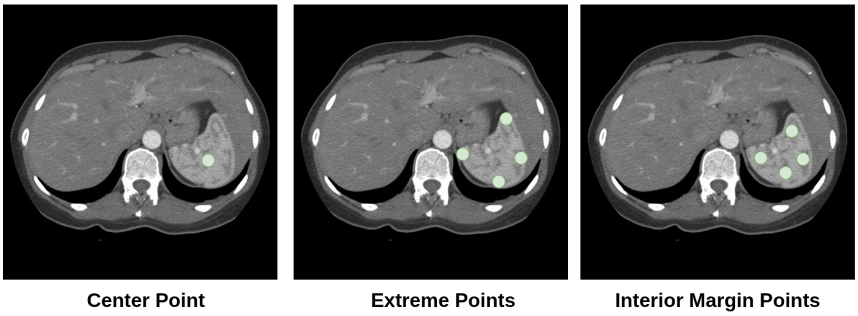


Figure 22: Rule-based click simulation. Extreme points place clicks on the left-, right-, top-, and bottom-most coordinates. Interior margin points perturb extreme points toward the object's center.

margin points [38] as seen in Figure 22. These approaches can be defined as a deterministic mapping $f : \mathcal{L} \mapsto \mathcal{C}$ that always produces the same result as defined in Equation 13:

$$\mathcal{C} = f(\mathcal{L}), \text{ with } f : \{0,1\}^{H\times W\times K} \to \mathbb{N}^3 \text{ or } f : \{0,1\}^{H\times W} \to \mathbb{N}^2 \tag{13}$$

5.1.2.2 *Scribble-based Non-iterative Robot Users*

Sampling-based. Scribbles, denoted as $\mathcal{S}$, can also be simulated as fragmented sets of coordinates by sampling points along object boundaries or other regions, the same way as clicks in Equation 12.

Rule-based. However, scribbles are more often simulated as linear or continuous structures drawn within the object mask. Non-iterative robot users generate these by skeletonizing the ground-truth mask to produce thin centralized structures as in Equation 14. Skeleton-based scribble maps can be formally defined as the medial axis, which is the set of all points that are equidistant from at least two points p, p' on the object's boundary $\mathcal{L}_{\text{boundary}}$:

$$\mathcal{S} = \{v \mid \exists p \neq p', \text{EDT}(v, \mathcal{L}_{\text{boundary}}) = \|p - v\| = \|p' - v\|\} \tag{14}$$

where EDT is defined in Equation 7.

5.1.2.3 *Bounding Box-based Non-iterative Robot Users*

Bounding boxes are generated directly from the spatial extent of the object mask $\mathcal{L}$. These simulations are purely rule-based and non-iterative, as sampling or iterative correction is generally not meaningful for bounding boxes. Given $\mathcal{L}$, the simulated bounding box $\mathcal{B}$ is defined by the minimum and maximum coordinates $(b_{\text{min}}^{(d)}, b_{\text{max}}^{(d)})$ of the object along each dimension d:

$$\mathcal{B} = \{\, (b_{\text{min}}^{(d)}, b_{\text{max}}^{(d)}) \mid \forall\, d \in \{1, \dots, D\}\}, \quad D \in \{2,3\}. \tag{15}$$

where $x^{(d)}$ is the d-th element of a vector x and b_{min} and b_{max} are the two opposing corner points that minimally enclose the ground-truth mask $\mathcal{L}$.

5.1.2.4 *Polygon-based Non-iterative Robot Users*

Polygons $\mathcal{P}$ are simulated by subsampling object boundaries to produce a discrete set of points along the boundary:

$$\mathcal{P} = \{p \mid p \sim P(\mathcal{L}_{\text{boundary}})\}, p \in \mathbb{N}^D, D \in \{2, 3\} \tag{16}$$

where $\mathcal{L}_{\text{boundary}}$ is the object boundary, and P is a distribution controlling how vertices are sampled. Rule-based robot users for polygon vertices do not exist to the best of our knowledge.

5.1.3 *Iterative Robot Users*

Iterative simulation methods emulate the loop of human interactions during model deployment, where an annotator repeatedly corrects the model's predictions in a human-in-the-loop scenario. In these simulations, each subsequent interaction is either sampled from missegmented regions or determined by rules, such as selecting the center of the largest error.

Methodologically, the interaction types and simulation strategies are the same as those used by non-iterative robot users, as described in Equations 12-16. The key difference is that the first interactions are typically simulated using the ground-truth mask $\mathcal{L}$, while subsequent interactions are guided by the model's error map $\mathcal{E}$. For example, a sampling-based iterative robot user may initially sample clicks from the ground-truth mask and then sample future clicks from the model's error regions. Iterative robot users can also leverage the model's uncertainty as an auxiliary signal, placing interactions in areas where the model is less confident.

5.1.4 *Robot Users During Training and Evaluation*

Interaction simulation can be applied during training, evaluation, or both.

Training. Simulating interactions during training is common, as collecting real interactions for large datasets is time-consuming and expensive, particularly when models are trained over multiple epochs. Performing ablation studies would further multiply the required interactions.

There are, however, a few exceptions. In *active learning* scenarios, models are first trained on a small labeled subset and then applied to unlabeled data. The most informative samples are identified based on model predictions, and annotators provide interactions for these samples to fine-tune the model. This iterative process continues until the annotators are satisfied with the model's performance [37, 62, 70, 92]. Another exception is *online learning*, where the model is updated in real-time through interactions with a human annotator. In such setups, interaction simulation during training is unnecessary [19, 65, 85]. Both the active learning and online learning paradigms are discussed in much greater detail in our taxonomy in Chapter 6.

Evaluation. Interactive models can be evaluated in two main ways. One approach is to conduct a real user study with medical annotators, who provide feedback to the model and perform interactive annotation for a specific task. This type of evaluation gives a realistic assessment of model performance for the target user group. However, such studies are typically limited in both the number of participants and the number of images, due to the high cost of medical data annotation and the availability of qualified personnel.

Consequently, most interactive models are evaluated using simulated interactions. Simulation allows for evaluation on a much larger number of images and supports exhaustive ablation studies, investigating how different model parameters affect performance (as in our study on guidance signals involving 100 models from Chapter 3).

Because real user studies are constrained to small sample sizes, and large-scale evaluation is only feasible through simulation, it is essential to implement realistic robot users. In our experiments with *SW-FastEdit* in Chapter 4, we observed that traditional robot users consistently outperformed real annotators. In the next section, we analyze the reasons behind this discrepancy and propose a more realistic robot user to mitigate the simulated-to-real user gap.

5.2 REALISTIC CLICK SIMULATION FOR PET/CT INTERACTIVE LESION SEGMENTATION

This section is based on our publication in *MICCAIw 2024* [285][2].

Having defined the various types of robot users used in literature in Section 5.1, a natural question arises:

How realistic are existing robot users in practice?

Answering this question requires direct comparison between simulated robot users and real medical annotators. Such comparisons are essential during evaluation, since the vast majority of interactive segmentation models[3] (over 80%) report performance obtained exclusively with simulated users [209]. If these simulations fail to reflect realistic annotation behavior, the reported results may not reflect the true model performance when deployed in clinical practice.

To investigate this, we conduct a user study with four medical annotators and observe that existing robot users differ substantially from real users in both segmentation performance and interaction behavior. We refer to this discrepancy as the *user shift*. To quantify the simulated-to-real user shift, we introduce five evaluation metrics **(M1)–(M5)**. Based on the findings from this initial study, we design a more realistic robot user that incorporates human factors such as click variability and inter-annotator disagreement. A second user study with four additional annotators confirms that our proposed robot user consistently reduces the simulated-to-real user shift compared to traditional ones. Moreover, we show that our user shift metrics **(M1)–(M5)** strongly correlate with actual segmentation performance differences, indicating that larger shifts correspond to greater discrepancies in Dice scores. This also results in much more realistic Dice scores produced when applying our robot user for model evaluation. By employing

[2] ADSMI@MICCAI 2024: MICCAI Workshop on Advancing Data Solutions in Medical Imaging AI: https://adsmi-miccai.github.io/

[3] Deep learning–based medical interactive segmentation models published between 2016–2023.

our robot user, large-scale and cost-efficient evaluations of interactive segmentation models become possible, without sacrificing the realism and fidelity of real user studies.

Our work highlights the importance of realistic simulation of human interaction and contributes to the field of medical interactive segmentation through the following:

1. We evaluate four existing robot users **(R1)–(R4)** on the *autoPET* dataset [279] and perform two user studies, each involving four medical annotators, to characterize the divergence between simulated and real user performance.

2. We introduce five evaluation metrics **(M1)–(M5)** that quantify the simulated-to-real user shift in terms of interaction behavior and agreement with ground-truth annotations.

3. We propose a novel robot user that addresses the shortcomings identified in (1) by simulating interactions that strategically deviate from the ground-truth labels, inspired by the large inter-annotator disagreement. This approach reduces both the measured user shift (as defined in 2.) and the segmentation performance gap to real users across both studies.

5.2.1 *Experimental Setup*

5.2.1.1 *Dataset, Interactive Segmentation Model, and Robot Users*

Dataset and Model Architecture. We use the pre-trained *SW-FastEdit* [239] interactive model based on *MONAI Label* [149] with a U-Net backbone [127] and conduct our user studies on the openly available *autoPET* [279] dataset, which consists of 1014 PET/CT volumes with annotated lesions of melanoma, lung cancer, or lymphoma. We exclusively utilize PET data and use *SW-FastEdit*'s test split, as specified in Chapter 4. The PET volumes have a voxel size of $2.0 \times 2.0 \times 3.0\text{mm}^3$ and an average resolution of $400 \times 400 \times 352$ voxels.

Iterative Robot Users. We explore iterative robot users that simulate clicks in a loop of 10 iterations. In each click iteration $i \in \{1, ..., 10\}$, a robot user R simulates a click, denoted as $\texttt{clicks}(R, I)[i] \in \mathbb{N}^3$, and combines it with the image $I \in \mathbb{R}^{W \times H \times K}$ as a joint input, where $W \times H \times K$ are the image dimensions. Using this joint input, the model predicts a segmentation mask $\texttt{pred}(I)[i] \in \{0, 1\}^{W \times H \times K}$. Then, the missegmented regions within this prediction, denoted as $\texttt{err}(I)[i] \in \{0, 1\}^{W \times H \times K}$, are used to generate $\texttt{clicks}(R, I)[i + 1]$ for the next iteration, forming an interaction loop.

(R1) Center Click: A common approach is to simulate clicks in the center of the largest missegmented component [64, 209]. However, the first click is placed in the center of the largest component of the label $\mathcal{L}$. This is defined as:

$$\texttt{clicks}(R1, I)[i] = \begin{cases} \texttt{center}(\texttt{largest\_component}(\mathcal{L})), & \text{if } i = 1 \\ \texttt{center}(\texttt{largest\_component}(\texttt{err}(I)[i-1])), & \text{if } i > 1 \end{cases} \tag{17}$$

where $\mathcal{L} \in \{0,1\}^{W \times H \times K}$ is the ground-truth label for image I, $\texttt{center}(\cdot)$ computes the geometric center of a component as in [64], and $\texttt{largest\_component}(\cdot)$ computes the largest connected component.

(R2) Uncertainty Sampling: Following Zheng et al. [48], we sample a click in each iteration using the epistemic uncertainty of the model as a sampling distribution, defined as:

$$\texttt{clicks}(R2, I)[i] \sim \begin{cases} \texttt{uniform}(\mathcal{L}), & \text{if } i = 1 \\ \texttt{epistemic}(\texttt{pred}(I)[i-1]), & \text{if } i > 1 \end{cases} \tag{18}$$

where $\texttt{epistemic}(\cdot)$ is the normalized epistemic uncertainty in $[0,1]$ using Monte Carlo Dropout [287], and $\texttt{uniform}(X)$ defines a uniform distribution over the non-zero entries of X.

(R3) Euclidean Distance Transform (EDT) Sampling: Many prior methods [10, 52, 67, 82, 239] apply the EDT on missegmented regions as a sampling distribution for clicks:

$$\texttt{clicks}(R3, I)[i] \sim \begin{cases} \texttt{uniform}(\mathcal{L}), & \text{if } i = 1 \\ 1 - \texttt{EDT}(\texttt{err}(I)[i-1]), & \text{if } i > 1 \end{cases} \tag{19}$$

where $\texttt{EDT}(\texttt{err}(I)[i-1])$ is the EDT computed over the missegmented regions $\texttt{err}(I)[i-1]$ from the previous iteration.

(R4) Uniform Sampling: The final robot user samples uniformly either from the previous error [288] or from the label $\mathcal{L}$ for the first click as:

$$\texttt{clicks}(R4, I)[i] \sim \begin{cases} \texttt{uniform}(\mathcal{L}), & \text{if } i = 1 \\ \texttt{uniform}(\texttt{err}(I)[i-1]), & \text{if } i > 1 \end{cases} \tag{20}$$

Note: In each iteration, we simulate two types of clicks - a lesion and a background click. We designate the under- and oversegmented regions as missegmented areas $\texttt{err}(I)[i]$ for the "lesion" and "background" classes, respectively, and simulate a click for each class independently. We also omit the class labels ("lesion" and "background") in Equations 17-20, for clarity.

5.2.1.2 *Evaluation Metrics*

For all evaluation metrics, we denote Ω as the set of PET images labeled in a user study, $\mathcal{A}$ as the set of real annotators participating in the study, and the fixed number of clicks per image as $N = 10$. We visualize examples for **(M1)–(M4)** in Figure 23.

(M1) The Label Agreement for an annotator $A \in \mathcal{A}$ is defined as:

$$\mathbf{M}_1(A) = \frac{1}{|\Omega|} \frac{1}{10} \sum_{I \in \Omega} \sum_{i=1}^{10} [\mathcal{L}[\texttt{clicks}(A, I)[i]] = 1] \tag{21}$$

where $[\cdot]$ is the Iverson bracket. **(M1)** measures to what extent an annotator's clicks agree with the ground-truth labels of the PET images, i.e., an approximation of the inter-annotator agreement between annotator A and the annotators that have originally labeled the dataset. It is essentially the proportion of annotator's A clicks that are within the ground-truth labels.

(M2) The Centerness for an annotator $A \in \mathcal{A}$ is defined as:

$$\mathbf{M}_2(A) = \frac{1}{|\Omega|} \frac{1}{|\bar{C}(A,I)|} \sum_{I \in \Omega} \sum_{c \in \bar{C}(A,I)} \frac{\texttt{bound}(c,\mathcal{L})}{\texttt{bound}(c,\mathcal{L}) + \texttt{cent\_dist}(c,\mathcal{L})} \tag{22}$$

where $\bar{C}(A,I) = \{c \mid c \in \texttt{clicks}(A,I) \text{ and } \mathcal{L}[c] = 1\}$ is the set of clicks of annotator A that agree with the ground-truth label $\mathcal{L}$, $\texttt{bound}(c,\mathcal{L})$ is the minimum distance of click c to the boundary of the label $\mathcal{L}$, and $\texttt{cent\_dist}(c,\mathcal{L})$ is the minimum distance of click c to the center of the label $\mathcal{L}$. Small **(M2)** values indicate that clicks that agree with the label $\mathcal{L}$ are placed near the boundary, whereas large values of **(M2)** mean that clicks are placed near the central regions of the label. In short, this metric quantifies whether annotator A is more likely to click in central regions or near the boundary.

(M3) The Click Diversity for annotator A is defined as:

$$\mathbf{M}_3(A) = \frac{1}{|\Omega|} \sum_{I \in \Omega} \frac{|\{\widetilde{\mathcal{L}} \mid \widetilde{\mathcal{L}} \in \texttt{components}(\mathcal{L}) \text{ and } \exists c \in \texttt{clicks}(A,I) : \ c \in \widetilde{\mathcal{L}}\}|}{\min(|\texttt{components}(\mathcal{L})|, |\texttt{clicks}(A,I)|)} \tag{23}$$

where $\texttt{components}(\cdot)$ is the set of all connected components. This metric measures to what extent clicks are spread out in different connected components $\widetilde{\mathcal{L}}$ in the label $\mathcal{L}$.

(M4) The Label Proximity for an annotator A is defined as:

$$\mathbf{M}_4(A) = \frac{1}{|\Omega|} \frac{1}{|\hat{C}(A,I)|} \sum_{I \in \Omega} \sum_{c \in \hat{C}(A,I)} \frac{1}{d(c,\mathcal{L})} \tag{24}$$

where $\hat{C}(A,I) = \{c \mid c \in \texttt{clicks}(A,I) \text{ and } \mathcal{L}[c] = 0\}$ is the set of clicks of annotator A that do not agree with label $\mathcal{L}$, and $d(c,\mathcal{L}) = \min(\{\|c-p\| \mid p \in \mathbb{N}^{W \times H \times K} \text{ and } \mathcal{L}[p] = 1\})$ is the minimum distance of the annotator's click to the label. **(M4)** computes the average inverse distance of the annotator clicks that do not agree with the ground-truth label to the label $\mathcal{L}$. Higher **(M4)** values suggest disagreeing clicks are close to the label boundary, e.g., because the object's edges are ambiguous. However, lower values of **(M4)** indicate that disagreeing clicks are far from any component of the label $\mathcal{L}$, suggesting a systematic disagreement with the original annotations, e.g., entire unannotated regions from the perspective of A.

(M5) The Consistent Improvement is defined in [91] and in Chapter 3 as:

$$\mathbf{M}_5(A) = \frac{1}{|\Omega|} \frac{1}{10} \sum_{I \in \Omega} \sum_{i=1}^{10} [\texttt{dice}(A,I)[i] > \texttt{dice}(A,I)[i-1]] \tag{25}$$

where $\texttt{dice}(A,I)[i]$ is the Dice score after annotator A's i^{th} click on image I.

(M6) The User Shift determines the mean absolute difference in all metrics **(M1)-(M5)** between a simulated robot user R and all real annotators $\mathcal{A}$:

$$\mathbf{M}_6(R, \mathcal{A}) = \frac{1}{|\mathcal{A}|}\frac{1}{5}\sum_{A \in \mathcal{A}}\sum_{i=1}^{5} |\mathbf{M}_i(R) - \mathbf{M}_i(A)| \tag{26}$$

(M7) The Dice Difference for a robot user R to the real annotators $\mathcal{A}$ is defined as:

$$\mathbf{M}_7(R, \mathcal{A}) = \frac{1}{|\Omega|}\frac{1}{|\mathcal{A}|}\frac{1}{10}\sum_{I \in \Omega}\sum_{A \in \mathcal{A}}\sum_{i=1}^{10} |\mathtt{dice}(A, I)[i] - \mathtt{dice}(R, I)[i]| \tag{27}$$

Metric **(M6)** quantifies the fidelity of the robot user in emulating annotator behavior, while **(M7)** evaluates its ability to reproduce the segmentation performance of the interactive model as used by real annotators.

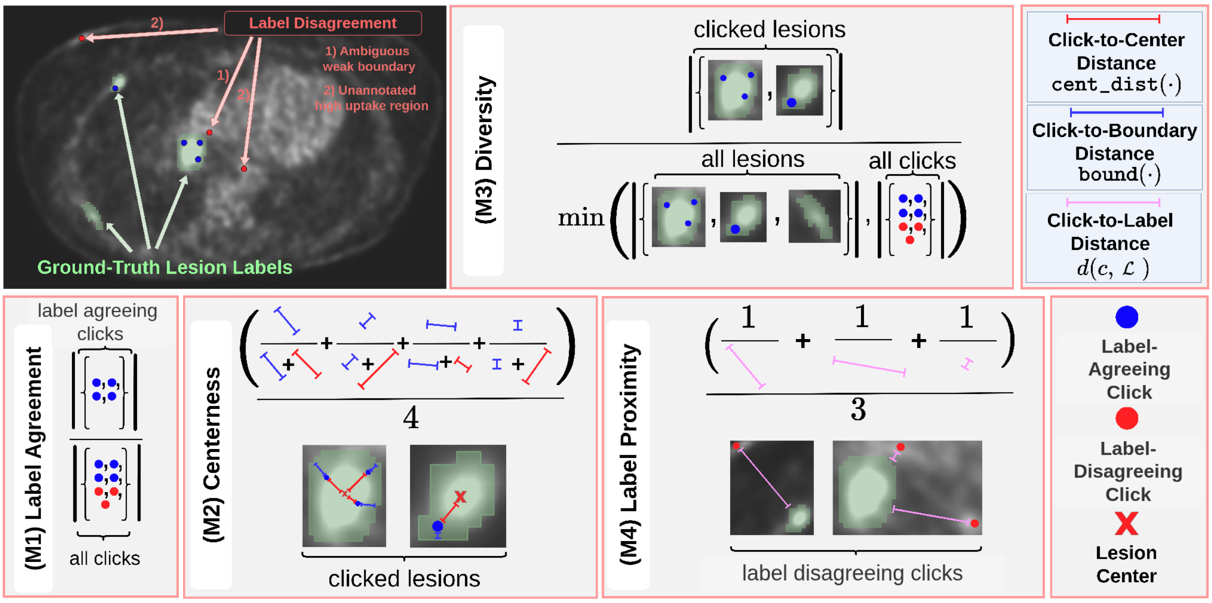


Figure 23: Visual depictions of our proposed evaluation metrics **(M1)–(M4)** and examples of label disagreement (top left). The blue and red dots are an annotator's clicks.

5.2.1.3 *User Studies*

We conducted two user studies, organized into **Phase 1** and **Phase 2**, each involving four annotators with medical backgrounds. In both studies, annotators were instructed to place **10 lesion** and **10 background** clicks per volume, updating the model prediction after each pair of clicks. This procedure replicates the workflow of the robot users and enables direct comparison between human and simulated interactions.

Phase 1 — Analysis and Robot User Design. Phase 1 aimed to investigate how existing robot users compare to real annotator interactions and to develop an improved robot user model. The following steps were conducted:

1. Four medical annotators independently labeled the same **10 PET volumes** from the *SW-FastEdit* test split, performing 10 lesion and 10 background clicks in an iterative loop.
2. We simulated corresponding click sequences for existing robot users **(R1)–(R4)**.
3. We evaluated the robot users **(R1)–(R4)** across all metrics **(M1)–(M7)**.
4. Observing a significant user shift for all existing robot users **(R1)–(R4)**, we designed a new robot user **(R5)**.
5. Using the results of Phase 1, we optimized the hyperparameters p_{perturb} and p_{system} of **(R5)** to minimize the user shift metric **(M6)**.
6. Finally, we evaluated the Dice difference **(M7)** for our robot user with its optimal hyperparameters.

Phase 2 — Validation. Phase 2 served as an independent validation study to ensure that the optimized robot user **(R5)** generalizes beyond the data and annotators used in Phase 1. The following steps were conducted:

1. A new set of four medical annotators labeled **6 additional PET volumes**.
2. We evaluated the user shift **(M6)** and Dice difference **(M7)** for all robot users **(R1)–(R5)** with respect to the new annotators and volumes.

Both user studies were done in collaboration with the Annotation Lab Essen (`https://annotationlab.ikim.nrw/`).

5.2.2 *Results: Phase 1*

We evaluated *SW-FastEdit* using all robot users **(R1)—(R4)** as well as the real user interactions collected from the user study. The resulting Dice curves are shown in Figure 24. A clear performance gap can be observed between the robot users and real annotators (black curve). This suggests that current robot users tend to overestimate real-world performance, even when using the same model and data. In practice, this leads to overly optimistic expectations about model effectiveness in clinical applications.

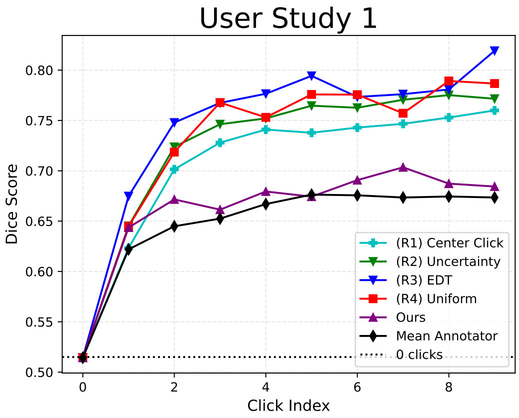


Figure 24: Dice curves of robot users for Phase 1.

Table 10 shows that the user shift is substantial, with robot users differing on average by **10%** Dice per click from the real performance observed in the user study. To better understand this discrepancy, we analyzed the clicks made by human annotators during the study. We found that **25% of the clicks** were located outside the ground-truth labels, indicating a significant disagreement with the reference annotations. This disagreement manifested in two main ways, illustrated in Figure 23 on the top left: (1) clicks placed just outside the ground-truth regions but close to the annotated lesions, likely due to weak lesion boundaries in the PET images; and (2) clicks placed in entirely new, previously unannotated high-uptake regions. These two types of labeling discrepancies motivated the development of a new robot user, **(R5)**, specifically designed to capture this behavior and better reflect realistic annotation patterns.

		(R1)	**(R2)**	**(R3)**	**(R4)**
User Study 1	**(M6)** User Shift	27.4	35.0	28.5	29.5
	(M7) Dice Difference	8.7	10.0	9.2	11.6

Table 10: User Shift **(M6)** and Dice Difference **(M7)** of existing robot users on the first user study.

(R5) Our Robot User: Based on insights from the first user study, where annotators occasionally placed clicks outside the ground-truth regions, we designed a new robot user to model this inter-annotator disagreement. To account for slight spatial inaccuracies near lesion boundaries, we introduce random click perturbations with probability p_{perturb}, allowing clicks to be placed outside the lesion. Additionally, to emulate clicks in unannotated high-uptake regions, we incorporate systematic label non-conformity by uniformly sampling clicks outside the ground-truth labels with probability p_{system}. Our robot user is formally defined as:

$$\texttt{clicks}(R5, I)[i] = \begin{cases} \texttt{clicks}(R1, I)[i] & \text{if } p_{i,1} \geqslant p_{\text{perturb}} \text{ and } p_{i,2} \geqslant p_{\text{system}} \\ \texttt{clicks}(R1, I)[i] + \tilde{z}, & \text{if } p_{i,1} < p_{\text{perturb}} \text{ and } p_{i,2} \geqslant p_{\text{system}} \\ \tilde{s}, & \text{if } p_{i,1} \geqslant p_{\text{perturb}} \text{ and } p_{i,2} < p_{\text{system}} \\ \tilde{s} + \tilde{z}, & \text{if } p_{i,1} < p_{\text{perturb}} \text{ and } p_{i,2} < p_{\text{system}} \end{cases} \tag{28}$$

where $\tilde{s} \sim \texttt{SUV}(I, \mathcal{L})$ and $\tilde{z} \sim \mathcal{U}_{[-a,a]^3}$. Here, $\texttt{SUV}(I, \mathcal{L})$ denotes a uniform distribution over the Standardized Uptake Values (SUVs) in I that are **outside** the label region $\mathcal{L}$ and **greater than the 95th percentile** of the SUV values in I. The variable $\tilde{z}$ represents a random perturbation vector with a maximum amplitude $a \in \mathbb{N}$. For each click, the parameters $p_{i,1}$ and $p_{i,2}$ are independently sampled from $\mathcal{U}_{[0,1]}$ to determine which of the four cases is applied.

Determining the Optimal Hyperparameters for (R5). To minimize the user shift **(M6)** for our robot user, we systematically explored combinations of p_{perturb}, p_{system}, and the perturbation amplitude a, evaluating each configuration using **(M6)**. The results of this hyperparameter search are shown in Figure 25. Spatial perturbations with $p_{\text{perturb}} \leqslant 75\%$ consistently outperform existing robot users in terms of user shift. The optimal configuration is achieved for $p_{\text{perturb}} = 25\%$ and $a = 35$, with performance degrading for $a > 35$ or $p_{\text{perturb}} = 100\%$ due to excessive spatial noise. Introducing systematic non-conformity further reduces user shift, with an optimal $p_{\text{system}} = 25\%$, mirroring the best perturbation probability. As both parameters achieve optimal performance at 25%, we also examined a mixed configuration using a joint probability of 25%. As shown in Table 11, this combination yields the lowest user shift and Dice difference, resulting in the best overall performance when $p_{\text{system}} = p_{\text{perturb}}$.

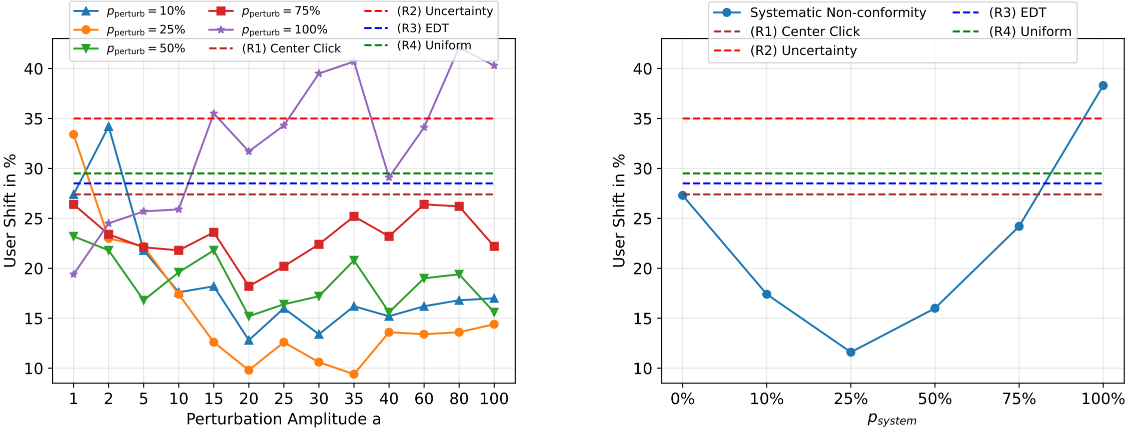


Figure 25: Analysis of p_{perturb} (left) and p_{system} (right) in our first user study.

		p_{perturb}	25%	19.6%	13.4%	6.7%	0%
		p_{system}	0%	6.7%	13.4%	19.6%	25%
User Study 1	**(M6)** User Shift		9.4	8.4	**6.8**	9.0	11.6
	(M7) Dice Difference		6.0	5.3	**3.6**	5.8	6.9

Table 11: User Shift **(M6)** and Dice Difference **(M7)** of our robot user **(R5)** on the first user study.

The results from our hyperparameter search show that incorporating a 25% label disagreement distributed equally among p_{perturb} and p_{system} aligns well with the 25% disagreement we observed with real annotators. We set $p_{\text{perturb}} = p_{\text{system}} = 13.4\%$ and $\alpha = 35$ as the optimal parameters for the second user study in Phase 2. We also visualize the Dice curve of our optimal **(R5)** in Figure 24 (purple line). The purple line aligns much better with what we see in practice when real annotators use *SW-FastEdit* (black line).

User Shift vs. Dice Difference. As the user shift **(M6)** only quantifies the behavioral shift, we examine its correlation with the Dice difference **(M7)** for all our robot user configurations in the first user study. Figure 26 reveals a Pearson correlation of $\rho = 0.89$ between the user shift and the Dice difference. Importantly, omitting any of our metrics **(M1)-(M5)** from **(M6)** decreases the correlation to $\rho < 0.8$. This confirms that our proposed metrics not only quantify the annotation style but also quantify how this style influences the segmentation performance.

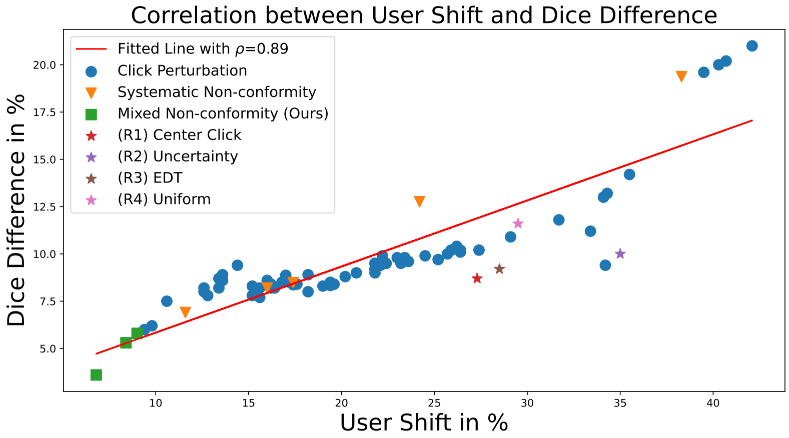


Figure 26: Correlation between **(M6)** and **(M7)** on the first user study results.

5.2.3 *Results: Phase 2*

The results from the second user study are presented in Table 12 and confirm that our robot user produces much more realistic interactions, both in terms of the user shift **(M6)** and the difference in Dice score from the real users **(M7)**. This is also evident in Figure 27, where the Dice curve of our robot user closely resembles that of the real medical annotators. The individual metrics **(M1)–(M5)** for both user studies are illustrated in Figure 28. Across all metrics, our robot user (cyan) provides a better representation of the average annotator (green) than all other robot users.

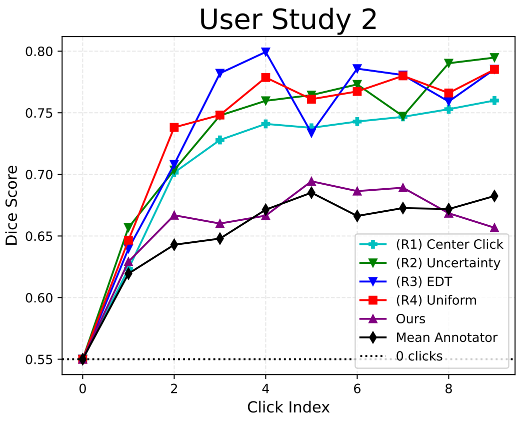


Figure 27: Dice curves of robot users for Phase 2.

		(R1)	(R2)	(R3)	(R4)	(R5)
User Study 1	**(M6)** User Shift	27.4	35.0	28.5	29.5	**6.8**
	(M7) Dice Difference	8.7	10.0	9.2	11.6	**3.6**
User Study 2	**(M6)** User Shift	30.0	31.7	33.8	30.0	**6.7**
	(M7) Dice Difference	8.5	9.0	7.0	7.5	**3.7**

Table 12: User Shift **(M6)** and Dice Difference **(M7)** of all robot users in both user studies.

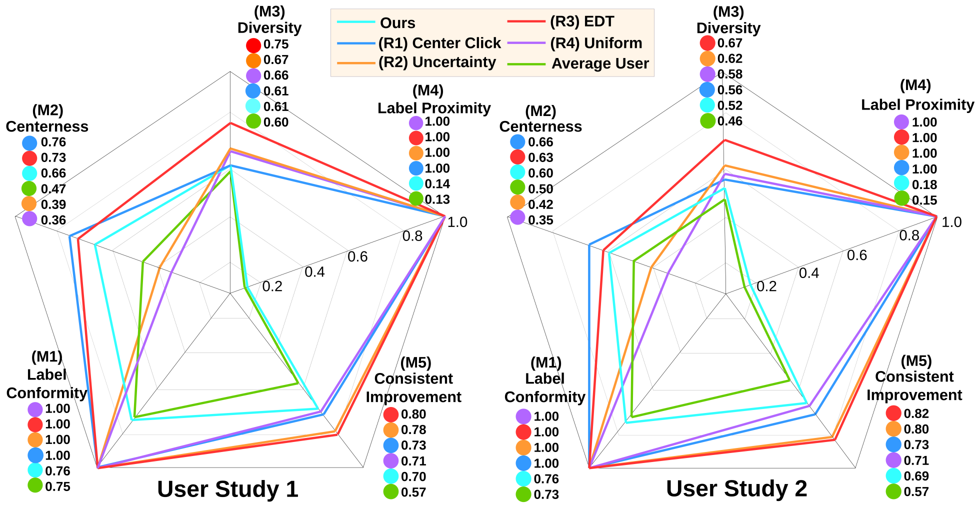


Figure 28: Metric values **(M1)-(M5)** of all robot users on both user studies.

Summary of results: Existing robot users **(R1)–(R4)** were found to produce overly optimistic results that fail to reflect the performance of real human annotators. This discrepancy was confirmed in two independent user studies. In the first study, we observed that about 25% of annotator clicks lay outside the ground-truth labels, largely due to ambiguous lesion boundaries and high-uptake regions not included in the annotations. These findings motivated the design of our new robot user **(R5)** that intentionally introduces such "disagreeing" clicks to better emulate realistic annotation behavior. The proposed approach yielded results that aligned more closely with real user performance and generalized well in a second user study involving new participants and unseen data.
Limitations: Our experiments are preliminary and limited to click-based interactions in whole-body PET lesion segmentation. The proposed robot user depends on two hyperparameters requiring careful tuning. Nonetheless, the broader insight remains that user interactions should not be assumed to perfectly conform to the original labels.

5.3 CHAPTER CONCLUSION

This chapter addressed the lack of theoretical and practical foundations for simulating user interactions in interactive segmentation. We began by formally defining the concept of a *robot user* (Definition 2), distinguishing between iterative and non-iterative formulations, and unifying prior approaches within a common mathematical framework (Equations 12–16).

We then systematically compared existing robot users with real medical annotators on the challenging task of whole-body PET lesion segmentation through two independent user studies. These experiments revealed a fundamental limitation of traditional simulation methods: they tend to produce overly optimistic results that do not correspond to the behavior of human annotators in practice. In particular, our analysis of annotator interactions uncovered frequent "disagreeing" clicks, i.e., outside the ground-truth annotations, arising from ambiguous lesion boundaries or unannotated high-uptake regions.

To address these discrepancies, we introduced a novel robot user that explicitly models inter-annotator disagreement by incorporating realistic perturbations and context-dependent click behavior. This new formulation significantly improved the realism of simulated interactions and better matched the outcomes observed in real user studies.

Beyond its empirical performance, our work questions a core assumption underlying traditional robot users—that simulated interactions must perfectly conform to ground-truth labels. We demonstrated that this assumption does not hold in complex, noisy medical imaging tasks where even experts may disagree. By highlighting this gap, the chapter establishes both a conceptual and methodological foundation for more realistic, human-aligned simulation of interactive segmentation.

We summarize the scientific impact of this chapter in four key contributions:

Contribution 1: We provide the first formal definition of a robot user and its subtypes (iterative and non-iterative), establishing a theoretical foundation for interactive segmentation (*TPAMI 2024* [209]).

Contribution 2: We present a comparative analysis of existing click-based robot users and real medical annotators across two independent user studies. Our analysis reveals fundamental flaws in traditional robot users, which assume full inter-annotator agreement and therefore yield overly optimistic results in simulated evaluations (*MICCAIw 2024* [285]).

Contribution 3: We introduce a novel, more realistic robot user for whole-body PET lesion segmentation that reproduces human-like interaction behavior by modeling inter-annotator disagreement and click variability (*MICCAIw 2024* [285]).

Contribution 4: We propose an evaluation framework to quantify the simulated-to-real user shift, comparing robot users and real annotators across five complementary metrics that jointly assess annotation style, degree of label disagreement, and the underlying nature of this disagreement in whole-body PET lesion segmentation (*MICCAIw 2024* [285]).

Our contributions highlight the crucial role of robot user design in the evaluation of interactive segmentation models, emphasizing the need to explore alternative interaction types (e.g., scribbles, bounding boxes) and broader segmentation tasks. Our proposed theoretical framework enables researchers to systematically categorize prior work on interaction simulation and gain insights for implementing their simulation methods. While our proposed robot user produces results that more closely reflect real annotator behavior than traditional methods, it remains task-specific and limited to click-based interactions. Nevertheless, it exposes fundamental flaws in the prevailing assumption that all experts consistently agree on all tasks, thereby motivating a rethinking of how simulated interactions are modeled and evaluated.

5.4 PART II CONCLUSION

This marks the conclusion of Part ii of this thesis. Throughout the previous three chapters, we have addressed three fundamental questions:

Chapter 3: How should one choose a guidance signal for their interactive model?

Chapter 4: How can we design efficient interactive models?

Chapter 5: How can we realistically simulate human interactions?

By focusing on these questions, we have provided solutions and insights that extend beyond specific tasks or model architectures, advancing the interactive segmentation paradigm along three key dimensions: **representation**, **efficiency**, and **simulation**.

In the following Part iii, we move beyond individual dimensions and examine the field of interactive segmentation in its full scope. Specifically, in Chapter 6, we conduct a systematic review of the entire domain and propose a comprehensive taxonomy that categorizes any interactive model—independent of its interaction type, architecture, or application. Chapter 7 then explores how to build and strengthen the interactive segmentation community through international collaboration and the organization of shared segmentation competitions. Finally, Chapter 8 introduces a large-scale, standardized dataset that consolidates 166 open-licensed datasets across nine imaging modalities, complemented by unified interaction simulation and evaluation protocols.

The overarching goal of Part iii is to bring together the literature (Chapter 6), the community (Chapter 7), and the datasets (Chapter 8) into a coherent, standardized framework for the field. Hence, the central research question for the next part is:

How can we standardize the field of medical interactive segmentation?

Part III

STANDARDIZATION OF INTERACTIVE SEGMENTATION

6

TAXONOMY AND PITFALLS OF INTERACTIVE SEGMENTATION

This chapter broadens the scope of this thesis to a global perspective, presenting a comprehensive systematic review of the field of deep interactive segmentation for medical images. To understand the current landscape and its challenges, it is essential to examine prior work in depth. We therefore conduct an exhaustive review of all relevant studies published between 2016 and 2023, from which we develop a novel taxonomy to categorize existing and future interactive segmentation approaches. In Section 6.1, we describe our systematic review process, introduce our novel taxonomy tree, and demonstrate how it can be applied to classify emerging methods by answering a sequence of questions. Section 6.2 identifies key pitfalls observed in the literature, including the lack of standardized comparisons, the absence of user-centered evaluation metrics, and the scarcity of interactive benchmarks. Finally, Section 6.3 explores opportunities to address these pitfalls, informed by current research trends, new datasets, and emerging model architectures.

6.1 DEEP INTERACTIVE SEGMENTATION OF MEDICAL IMAGES: A SYSTEMATIC REVIEW AND TAXONOMY

This **entire chapter** is based on our publication in *TPAMI 2024* [209], © IEEE.

6.1.1 *Motivation and Rationale for a Systematic Review*

The field of interactive segmentation has evolved rapidly in recent years, particularly with the advent of deep learning. Despite this progress, existing reviews remain limited in scope. Earlier surveys either focus on classical interactive segmentation techniques rather than modern deep learning approaches [145, 155, 156, 180], or exclude the medical domain altogether [157].

Historically, interactive segmentation methods originated from classical approaches such as active contour models [153] and *Graph Cut* [131], which relied on low-level image cues, such as pixel intensity gradients, to separate foreground from background. These traditional techniques often required handcrafted features to capture higher-level semantic information about the object of interest [5, 65, 140] and demanded manual parameter tuning, which was frequently domain- or image-specific [131, 141, 153].

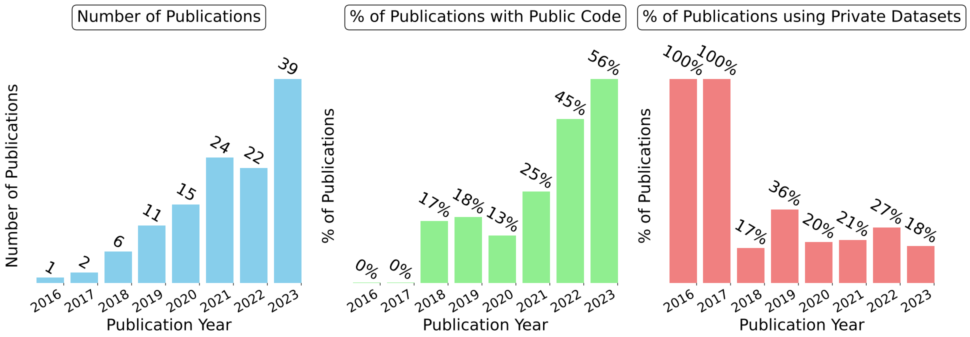


Figure 29: The growth of deep medical interactive segmentation in recent years (2016-2023).

Many of these limitations have been addressed by deep learning-based interactive segmentation methods, first popularized by Xu et al. [140]. These approaches leverage powerful end-to-end learned feature representations and to improve segmentation quality, robustness, and generalization.

However, no comprehensive review currently exists that focuses specifically on **deep learning-based** interactive segmentation in medical imaging, despite the field's rapid expansion, with more than 120 methods proposed between 2016 and 2023 (see Figure 29). The absence of a systematic overview not only hampers cumulative scientific progress by fostering redundant research but also makes it difficult for practitioners to identify the most suitable approaches for their clinical or research needs.

In this section, we address these shortcomings in a dedicated systematic review with the following key contributions:

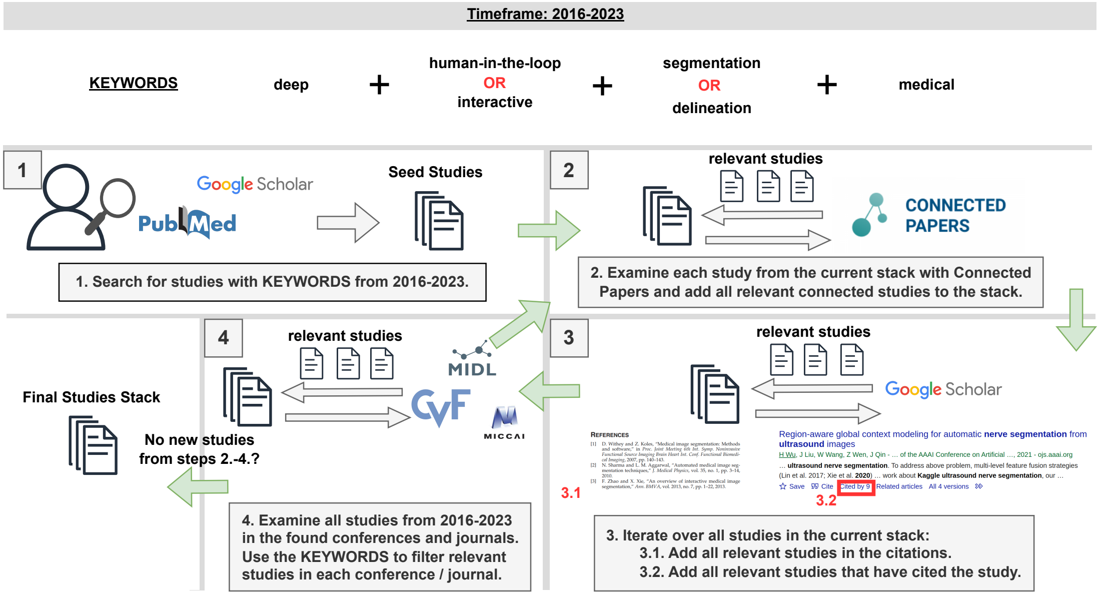
1. We provide a systematic review of 121 interactive segmentation methods in the medical domain following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [139].

2. We introduce a comprehensive taxonomy for deep interactive segmentation, allowing users to quickly comprehend the various existing approaches and select the best-fitting method for their task.

3. We perform an in-depth analysis of the current practices in the field, including prevalent datasets, segmentation targets, and validation metrics, as well as the adequacy of baselines and the reproducibility of prior work's results.

4. Based on this analysis, we provide a discussion of current challenges in the field and opportunities on how to effectively address them.

6.1.2 *Scope and Study Collection Strategy*



1. Search for studies with KEYWORDS from 2016-2023.

2. Examine each study from the current stack with Connected Papers and add all relevant connected studies to the stack.

REFERENCES

[1] D. Withey and Z. Koles, "Medical image segmentation: Methods and software," in *Proc. Joint Meeting 6th Int. Symp. Noninvasive Functional Source Imaging Brain Heart Int. Conf. Functional Biomedical Imaging*, 2007, pp. 140–143.

[2] N. Sharma and L. M. Aggarwal, "Automated medical image segmentation techniques," *J. Medical Physics*, vol. 35, no. 1, pp. 3–14, 2010.

[3] F. Zhao and X. Xie, "An overview of interactive medical image segmentation," *Ann. BMVA*, vol. 2013, no. 7, pp. 1–22, 2013.

Region-aware global context modeling for automatic **nerve segmentation** from **ultrasound** images
H Wu, J Liu, W Wang, Z Wen, J Qin - … of the AAAI Conference on Artificial …, 2021 - ojs.aaai.org
… **ultrasound nerve segmentation**. To address above problem, multi-level feature fusion strategies (Lin et al. 2017; Xie et al. **2020**) … work about **Kaggle ultrasound nerve segmentation**, our …
☆ Save Cite Cited by 9 Related articles All 4 versions

3. Iterate over all studies in the current stack:
 3.1. Add all relevant studies in the citations.
 3.2. Add all relevant studies that have cited the study.

4. Examine all studies from 2016-2023 in the found conferences and journals. Use the KEYWORDS to filter relevant studies in each conference / journal.

Figure 30: Search strategy in our systematic review for selecting relevant studies. The logos in steps 1 and 4 are illustrated only as examples for literature databases and venues, respectively. A full list of all venues (conferences, journals, and workshops) is given here[1].

We conduct a systematic review of deep learning-based interactive segmentation models applied in medical scenarios. Our review, being inherently technical in nature, aims to rigorously categorize and analyze relevant literature. Recognizing the need for a comprehensive reporting framework, we integrate as many components from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines by Moher et al. [139] as applicable to enhance the transparency and methodological clarity of our study. A detailed account of the adopted PRISMA components can be found in our PRISMA 2020 checklist tables[1].

We performed a literature search in several databases, including *PubMed*, *Google Scholar*, *IEEE Xplore*, *SpringerLink*, and *arXiv* (listed in Table 13), using specific keywords – [interactive], [human-in-the-loop], [segmentation], [delineation], [medical], and [deep]. The search was carried out on 31 July 2023, and we limited the publication period to cover the years 2016–2023 since the first deep learning interactive method originated in 2016 [140]. We removed duplicates, including pre-prints followed by their peer-reviewed versions. Subsequently, we conducted an initial manual screening of titles and abstracts to ensure that the selected studies are relevant. After this initial screening, full texts were retrieved and reviewed for eligibility based on specific inclusion criteria: 1) studies with English full texts; 2) studies that have undergone peer-review or have pre-prints submitted to the *arXiv* database; and 3) studies describing the

[1] https://ieeexplore.ieee.org/ielx8/34/10746266/10660300/supp1-3452629.pdf?arnumber=10660300

application of interactive segmentation models for a human medical purpose. Consequently, certain exclusion criteria were applied to maintain the focus and quality of the review: 1) studies lacking English full texts; 2) studies that utilize non-deep learning models; 3) studies that utilize interactive models solely on natural images; and 4) studies using medical images but not as the primary focus.

Literature Database	**Link**
Google Scholar	https://scholar.google.com/
PubMed	https://pubmed.ncbi.nlm.nih.gov/
IEEE Xplore	https://ieeexplore.ieee.org/Xplore/home.jsp
SpringerLink	https://link.springer.com/
arXiv	https://arxiv.org/

Table 13: List of all literature databases used in step 1 of our systematic search

We assessed a study's eligibility through a three-stage process: (1) we examine the title to decide if it focuses on deep medical interactive segmentation; (2) if the title is ambiguous, we read the abstract for confirmation; (3) in cases where the abstract remains unclear, we read the entire study to decide whether to include it in the review.

This search produced our initial *seed studies* stack as illustrated in Figure 30 (step 1). In addition to adhering to the PRISMA guidelines, we implemented three supplementary steps in our search strategy to maximize the retrieval of relevant studies and formed an iterative loop utilizing these steps. These steps are depicted as steps 2, 3, and 4 in Figure 30. In step 2, we incorporated the *Connected Papers* tool[2] to enhance our search process. This tool was applied to each of the already included studies from the *seed studies*, and we systematically screened all the suggested studies recommended by the tool, ensuring they met our predefined inclusion and exclusion criteria. In step 3, we manually inspected all the citations of each study in the *seed studies* and all of the studies that have cited this study using the "Cited by" function in Google Scholar. In step 4, we formed a list of all the peer-reviewed venues, and manually screened all of the publications from each venue in the timeframe 2016–2023 with our predefined keywords and added the relevant publications in our *seed studies*. We repeated steps 2, 3, and 4 and accumulated all relevant studies in our *seed studies* stack until no new relevant studies were found. Our search strategy found a total of 121 relevant publications.

After collecting all studies, we manually extracted from each study the following data items: 1) used imaging modalities; 2) used datasets along with provided links, if available; 3) prior interactive methods the study has compared to; 4) employed evaluation metrics; 5) type of interaction, e.g., clicks; 6) target structures for segmentation; 7) and, if applicable, a link to publicly available code. We cataloged all 121 reviewed studies and their data items in Tables 14, 15, and 16. This facilitates efficient navigation for future researchers seeking relevant interactive methods related to their own work.

[2] https://www.connectedpapers.com/

Paper	Year	Interaction	Guidance Signal	Target	Modality	Link
DeepCut [1]	2016	Bounding Box	Implicit	fetal brain, fetal lungs	MRI	Link
UI-Net [2]	2017	Scribbles	Identity	liver lesions	CT	Link
Sun et al.[3]	2017	Clicks	Location Prior	prostate	MRI	Link
Can et al.[4]	2018	Scribbles	Identity	cardiac structures, prostate	MRI	Link
DeepIGeoS [5]	2018	Scribbles	Geodesic Maps	placenta, brain tumors	MRI	Link
BIFSeg [6]	2018	Scribbles	Identity	placenta, kidneys, fetal brain+lungs, brain tumors	MRI	Link
InterCNN [7]	2018	Scribbles	Identity	prostate	MRI	Link
Dhara et al.[8]	2018	Scribbles	Identity	brain tumors	MRI	Link
Tang et al.[9]	2018	Bounding Box	Implicit	lung nodules, liver lesions, lymph nodes	CT	Link
Sakinis et al.[10]	2019	Clicks	Heatmaps	$\geqslant 10$ targets	CT	Link
Zhou et al.[11]	2019	Scribbles	Identity	brain tumors	MRI	Link
Khan et al.[12]	2019	Clicks	Chebychev Maps	heart, aorta, trachea, esophagus	CT	Link
DeepIGeoSv2 [13]	2019	Scribbles	Euclidean Maps	brain stem, parotid, optic nerve, optic chasm	CT	Link
iW-Net [14]	2019	Clicks	Attraction Field Map	lung nodules	CT	Link
Roth et al.[15]	2019	Clicks	Heatmaps	liver, spleen, prostate, cardiac structures	CT, MRI, US	Link
Ceronne et al.[16]	2019	Clicks	Disks	neuron cells	Microscopy	Link
Zheng et al.[17]	2019	Scribbles	Implicit	pancreas	CT	Link
Chao et al.[18]	2019	Scribbles	Implicit	esophageal cancer	PET/CT	Link
Längkvist et al.[19]	2019	Scribbles	Identity	lung structures	CT	Link
Wang et al.[20]	2019	Polygons	Graph	liver	CT	Link
Boers et al.[21]	2020	Scribbles	Implicit	pancreas	CT	Link
UGIR [22]	2020	Scribbles	Geodesic Maps	fetal brain	MRI	Link
IterMRL [23]	2020	Clicks	Geodesic Maps	brain tumors, cardiac structures, prostate	MRI	Link
Raju et al.[24]	2020	Clicks	Heatmaps	liver	CT	Link
BS-IRIS [25]	2020	Clicks	Geodesic Maps	brain tumors, cardiac structures, prostate	MRI	Link
NuClick [26]	2020	Clicks + Scribbles	Disks + Identity	intestinal glands, cell nuclei, white blood cells	Microscopy	Link
Kitrungrotsakul et al.[27]	2020	Scribbles	Identity	liver	CT	Link
IRIS [28]	2020	Clicks	Implicit	aorta	CTA	Link
Hu et al.[29]	2020	Clicks	Geodesic Maps	brain tumors	MRI, CT	Link
Tian et al.[30]	2020	Polygons	Graph	prostate, cardiac structures	MRI	Link
Chao et al. 2 [31]	2020	Scribbles	Implicit	nasopharangeal and esophageal cancer	PET/CT	Link
Tang et al. 2 [32]	2020	Clicks	Euclidean Maps + Disks	lung nodules, liver lesions, lymph nodes	CT	Link
Jinbo et al.[33]	2020	Scribbles	Heatmaps	liver	CT	Link
Girum et al.[34]	2020	Clicks	Implicit	prostate, cardiac structures	CT, US	Link
Ho et al.[35]	2020	Scribbles	Identity	osteosarcoma	Microscopy	Link
Foo et al.[36]	2021	Scribbles	Longest Axis Lines	COVID19 lesions	CT	Link
Menon et al.[37]	2021	Scribbles	Implicit	colorectal cancer, breast cancer	Microscopy	Link
MIDeepSeg [38]	2021	Clicks	Exp. Geodesic Maps	placenta, spleen, kidney, prostate, fetal brain	CT, MRI, US	Link
Feng et al.[39]	2021	Clicks	Disks	liver, kidney, stomach, breast	CT	Link
Roth et al. 2 [40]	2021	Clicks	Heatmaps	spleen, liver, pancreas, kidneys, gallbladder	CT	Link
Sambaturu et al.[41]	2021	Scribbles	Identity	$\geqslant 10$ targets	Microscopy, MRI, CT	Link
Zhou et al. 2 [42]	2021	Scribbles	Identity	lung, colon, kidney, kidney tumors	CT	Link
Williams et al.[43]	2021	Scribbles	B-splines	levator hiatus	US	Link
WDTISeg [44]	2021	Clicks	Euclidean/Geodesic Maps	breast cancer	US	Link
Li et al.[45]	2021	Clicks	Heatmaps	brain tumors, cardiac structures, spleen, liver	MRI, CT, CT/MRI	Link
Deng et al.[46]	2021	Scribbles	Implicit	aortic system, brain tumors	CTA, MRI	Link
Zhang et al.[47]	2021	Clicks	Implicit	kidney tumors, prostate	CT, MRI	Link

Table 14: (Part 1) List of studies and their attributes in our systematic review.

Paper	Year	Interaction	Guidance Signal	Target	Modality	Link
Zheng et al. 2 [48]	2021	Scribbles	Implicit	skin lesions	Dermoscopy	Link
DINs [49]	2021	Clicks	Heatmaps	neurofibromatosis type I	MRI	Link
Tian et al. 2 [50]	2021	Polygons	Graph	prostate	MRI	Link
Jiang et al.[51]	2021	Clicks	Heatmaps	skin lesions	Dermoscopy	Link
Bai et al.[52]	2021	Clicks	Heatmaps	$\geqslant$ 10 targets	CT, MRI	Link
Deepscribble [53]	2021	Scribbles	Euclidean Maps	liver tumors	Microscopy	Link
Attention-RefNet [54]	2021	Scribbles	Geodesic Maps	COVID19 lesions	CT	Link
Daulatabad et al.[55]	2021	Clicks	Disks	thyroid nodules	US	Link
Manh et al.[56]	2021	Clicks	Implicit	Z-line	Endoscopy	Link
Trimpl et al.[57]	2021	Scribbles	Identity	$\geqslant$ 10 targets	CT	Link
PiPo-Net [58]	2021	Polygons	Identity	breast cancer	Microscopy	Link
Jahanifar et al.[59]	2021	Scribbles	Identity	breast cancer	Microscopy	Link
Sun et al. 2 [60]	2022	Polygons	Implicit	prostate, nasoprahangeal cancer	MRI	Link
Shahedi et al.[61]	2022	Clicks	Disks	prostate	CT	Link
Atzeni et al.[62]	2022	Scribbles	Implicit	brain structures	MRI, Microscopy	Link
Bi et al.[63]	2022	Clicks	Euclidean Maps	skin lesions	Dermoscopy	Link
iSegFormer [64]	2022	Clicks	Disks	knee cartilage	MRI	Link
ECONet [65]	2022	Scribbles	Identity	COVID19 lesions	CT	Link
i3Deep [66]	2022	Scribbles	Identity	brain tumors, pancreas, COVID19 lesions	CT, MRI	Link
DeepEdit [67]	2022	Clicks	Heatmaps	prostate, prostate tumors	MRI, CT, CT/MRI	Link
Liu et al.[68]	2022	Clicks	Disks	lung, colon, pancreas	CT	Link
Shi et al.[69]	2022	Scribbles	Identity	colon cancer, lung cancer, kidney tumors, kidney	CT	Link
AnatomySketch [70]	2022	Scribbles + Polygons	Identity + Implicit	liver cancer, lung lobes, intervertebral disc	MRI, CT	Link
Galisot et al.[71]	2022	Bounding Box	Implicit	brain structures	MRI	Link
Lin et al. [72]	2022	Scribbles+Clicks+BBox	Heatmaps + Implicit	COVID19 lesions, brain tumors, brachial plexus, polyps, skin lesions	X-Ray, CT, MRI, US, Endoscopy, Dermoscopy	Link
Pirabaharan et al.[73]	2022	Clicks	Heatmaps	spleen, colon cancer	CT	Link
Mikhailov et al.[74]	2022	Clicks	Disks	uterus, bladder, uterine cavity, female pelvis tumors	MRI	Link
Pirabaharan et al. 2 [75]	2022	Clicks	Heatmaps	spleen, colon cancer	CT	Link
Chen et al.[76]	2022	Clicks	Heatmaps+Euclidean Maps	breast lesions	US	Link
Deep SED-Net [77]	2022	Scribbles	Identity	testicular cells	Microscopy	Link
Ju et al.[78]	2022	Clicks	Implicit	liver, kidneys, spleen	CT	Link
Ma et al.[79]	2022	Scribbles	Implicit	liver, spleen	CT, MRI	Link
Bai et al. 2 [80]	2022	Clicks	Heatmaps	nasopharangeal cancer	CT	Link
Zhou et al. 3 [81]	2023	Scribbles	Geodesic Maps	lung, colon, kidney, kidney tumors	CT, Endoscopy	Link
Hallitschke et al.[82]	2023	Scribbles	Geodesic Maps	lung cancer	PET/CT	Link
Liu et al. 2 [83]	2023	Clicks + Scribbles	Disks + Identity	$\geqslant$ 10 targets	CT	Link
Bruzadin et al.[84]	2023	Scribbles	Implicit	COVID19 lesions	CT	Link
Asad et al.[85]	2023	Scribbles	Identity	COVID19 lesions	CT	Link
Shahin et al.[86]	2023	Scribbles	Heatmaps	cardiac structures	US	Link
Zhuang et al.[87]	2023	Polygons	Graph	liver, spleen, kidneys	CT	Link
Ho et al. 2 [88]	2023	Scribbles	Identity	ovarian cancer	Microscopy	Link
Wei et al.[89]	2023	Scribbles	Identity	head-and-neck cancer	PET/CT/MRI	Link
Zhuang et al. 2 [90]	2023	Scribbles	Exp. Geodesic Maps	brain tumors, liver tumors	CT, MRI	Link
GtG [91]	2023	Clicks	Heatmaps	lung lesions, lymphoma, melanoma, spleen	CT, PET/CT	Link
Qu et al.[92]	2023	Scribbles	Identity	$\geqslant$ 10 targets	CT	Link

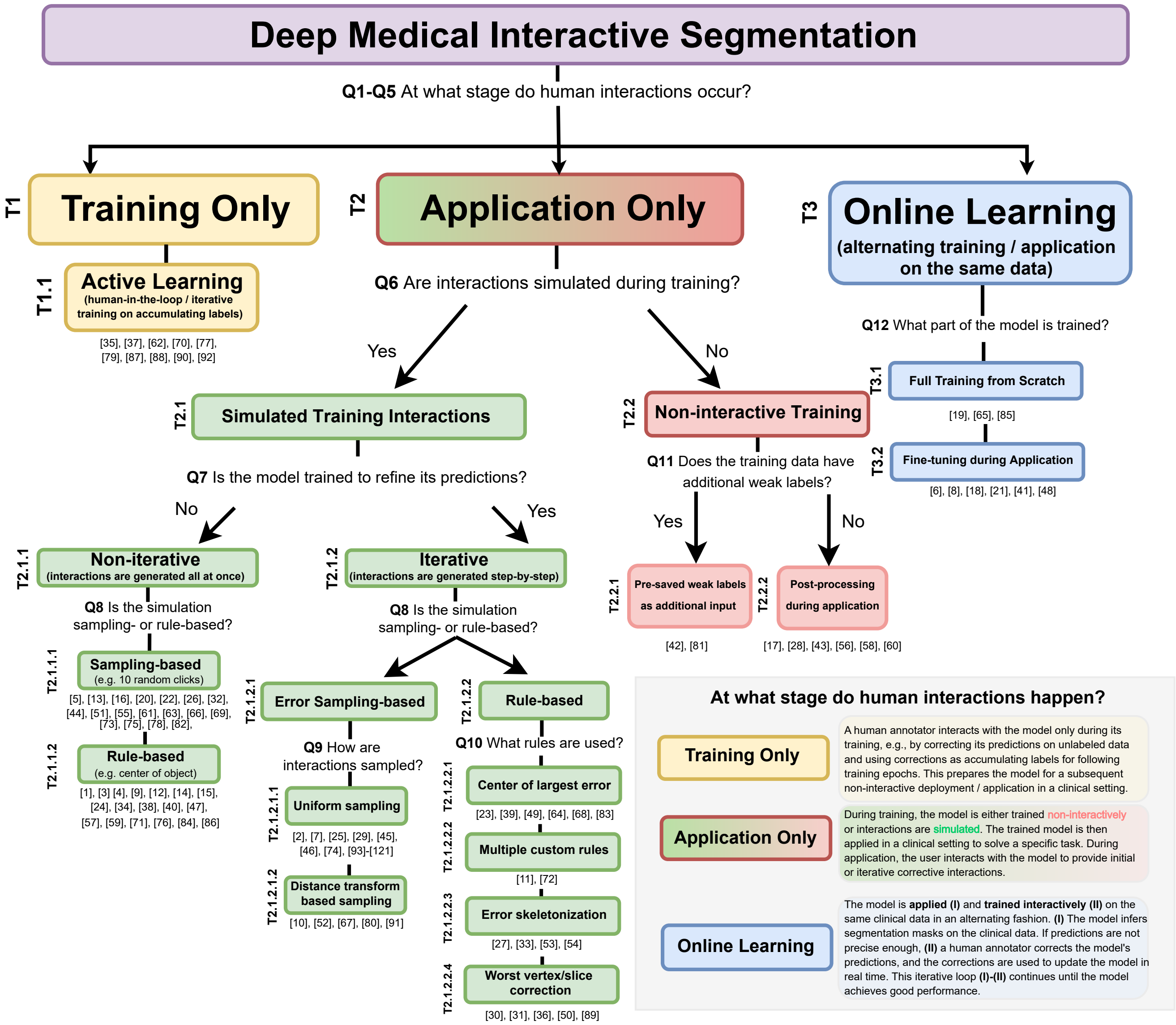
Table 15: (Part 2) List of studies and their attributes in our systematic review.

Paper	Year	Interaction	Guidance Signal	Target	Modality	Link
Mazurowski et al. [93]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	CT, MRI, US, X-Ray, PET/CT	Link
Deng et al. [94]	2023	Clicks+Bounding Box	Positional Encoding	cell nuclei, skin cancer	Microscopy	Link
SAM vs. BET [95]	2023	Clicks+Bounding Box	Positional Encoding	brain	MRI	Link
Putz et al. [96]	2023	Clicks+Bounding Box	Positional Encoding	brain tumors	MRI	Link
Hu et al. [97]	2023	Clicks+Bounding Box	Positional Encoding	liver tumors	CT	Link
SAM-Adapter [98]	2023	Clicks+Bounding Box	Positional Encoding	polyps	Endoscopy	Link
Medical SAM Adapter [99]	2023	Clicks	Positional Encoding	$\geqslant 10$ targets	CT, MRI, US, Fundus, Dermoscopy	Link
Ophthalmology SAM [100]	2023	Clicks+Bounding Box	Positional Encoding	blood vessels, retinal lesions	Fundus	Link
He et al. [101]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	MRI, US, CT, Endoscopy, Dermoscopy, X-Ray	Link
Shi et al. [102]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	Dermoscopy, Fundus, CT, MRI, Endoscopy, X-Ray, OCT	Link
GazeSAM [103]	2023	Eye Gaze	Positional Encoding	not specified	not specified	Link
SkinSAM [104]	2023	Bounding Box	Positional Encoding	skin lesions	Dermoscopy	Link
Wang et al. [105]	2023	Clicks+Bounding Box	Positional Encoding	surgical instruments	Endoscopy	Link
Cheng et al.. [106]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	US, Endoscopy, MRI, CT	Link
Mattjie et al. [107]	2023	Clicks+Bounding Box	Positional Encoding	skin lesions, lungs, femur, illium, polyps, breasts	X-Ray, US, Endoscopy, Dermoscopy	Link
Polyp-SAM [108]	2023	Bounding Box	Positional Encoding	polyps	Endoscopy	Link
PromptUNet [109]	2023	Clicks+Bounding Box + Scribbles	Positional Encoding + Identity	$\geqslant 10$ targets	CT, MRI, US, Fundus, Dermoscopy	Link
BreastSAM [110]	2023	Clicks+Bounding Box	Positional Encoding	breast cancer	US	Link
IAMSAM [111]	2023	Clicks+Bounding Box	Positional Encoding	breast cancer, colon, brain, prostate cancer,	Microscopy	Link
DeSAM [112]	2023	Clicks+Bounding Box	Positional Encoding	prostate	MRI	Link
Shen et al. [113]	2023	Clicks+Bounding Box	Positional Encoding	brain tumors	MRI	Link
Ning et al. [114]	2023	Clicks+Bounding Box	Positional Encoding	heart, thyroid, carotid artery	US	Link
Zhang et al. [115]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	CT	Link
MedLSAM [116]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	CT	Link
SAM-U [117]	2023	Bounding Box	Positional Encoding	optic disc, optic cup	Fundus	Link
3DSAM-adapter [118]	2023	Clicks	Positional Encoding	liver tumors, kidney tumors, pancreas tumors, colon cancer	CT	Link
Huang et al. [119]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	CT, MRI, Endoscopy, US, Fundus, Microscopy, Endoscopy, X-Ray	Link
MedSAM [120]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	CT, MRI, US, X-Ray, PET/CT, Microscopy, OCT, Endoscopy, Fundus	Link
SAM.MD [121]	2023	Clicks+Bounding Box	Positional Encoding	$\geqslant 10$ targets	CT	Link

Table 16: (Part 3) List of studies and their attributes in our systematic review.

6.1.3 *Taxonomy Tree*

After retrieving the 121 publications, we analyze the foundational principles of their methodologies and categorize them based on common characteristics. This procedure yields our proposed *taxonomy tree* in Figure 31, which functions as a navigational tool for existing medical interactive segmentation methods. This tool should help researchers categorize their approaches and steer them towards existing methods that align with their own. In this section, we provide detailed insights into the construction of our taxonomy.



A human annotator interacts with the model only during its training, e.g., by correcting its predictions on unlabeled data and using corrections as accumulating labels for following training epochs. This prepares the model for a subsequent non-interactive deployment / application in a clinical setting.

During training, the model is either trained non-interactively or interactions are simulated. The trained model is then applied in a clinical setting to solve a specific task. During application, the user interacts with the model to provide initial or iterative corrective interactions.

The model is **applied (I)** and **trained interactively (II)** on the same clinical data in an alternating fashion. **(I)** The model infers segmentation masks on the clinical data. If predictions are not precise enough, **(II)** a human annotator corrects the model's predictions, and the corrections are used to update the model in real time. This iterative loop **(I)-(II)** continues until the model achieves good performance.

Figure 31: Our proposed taxonomy tree for all the reviewed studies. The references for studies associated with a node are listed beneath the respective node.

We identified three paradigms ***T1-T3*** that are determined by the stage at which human interactions occur. These paradigms form the primary categorization in our taxonomy tree in Figure 31, and a summary of each paradigm can be found in the three boxes at the bottom right. We use the terms *training* and *application* as the building blocks of our taxonomy tree. In the training stage, the model undergoes optimization, where its weights are updated using a predetermined loss function. The subsequent application stage involves deploying the trained model on unseen data, utilizing its refined parameters to address specific clinical tasks. Depending on these two stages, interactions occur: 1) exclusively during application; 2) exclusively during training; 3) or in an alternating manner between both stages (online learning). These three paradigms constitute our proposed taxonomy and are described in detail in Sections 6.1.3.3, 6.1.3.4, and 6.1.3.5.

6.1.3.1 *Navigating our Taxonomy*

We facilitate the navigation in our taxonomy tree through decision guidelines, which pose specific questions *Q1-Q12* at each branching point focusing on the inherent strengths and weaknesses of each taxonomy node. By engaging with each question, users are encouraged to reflect on their objectives, resources, and specific use cases. With these questions integrated at every juncture, users can efficiently traverse our taxonomy, ultimately arriving at a category that aligns with their intended application.

The first juncture in our taxonomy tree categorizes methods based on where human interactions occur. We deem this decision as the most crucial in navigating our taxonomy and divide it into five questions *Q1-Q5* addressing: 1) availability of human interactors; 2) label availability; 3) model complexity; 4) model generalizability; and 5) number of training rounds. We depict the advantages and disadvantages of each node regarding these criteria in Figure 32 and present the questions in the following.

Q1. At which stage is an interactive user available?

Depending on the user availability, our three main taxonomy nodes present distinct challenges. In *training only* methods, users are required to correct model predictions for the most informative samples across multiple sessions, as models undergo iterative re-training after each correction session. Hence, users must be available at multiple points in time, but may correct the predictions at their own pace since the annotation process occurs offline. In *application only* methods, users utilize a pre-trained model and correct its predictions in real-time within one continuous interaction session or user study, demanding the users' undivided attention. In *online learning* methods, users must be available for both training and applying the model to the same data within a single uninterrupted session.

Recommendation. *Training only* methods are appropriate when users can participate in multiple sessions to annotate data at their own pace. In contrast, *application only* and *online learning* methods necessitate only one interactive round but require the user to be continuously available for the entire session. This is essential for measuring usability metrics such as the number of clicks or interaction time, or for training the interactive model in *online learning* methods.

Q2. How many annotations are available for the task?

The amount of available annotated training data is critical for an interactive model's development. However, some domains, like PET/CT, face annotation scarcity due to limited public datasets [154]. *Training only* methods begin with a small labeled fraction, termed the "starting budget," for initial pre-training of the model [62, 92]. The model then iteratively annotates the unlabeled portion across multiple rounds. In contrast, *application only* methods require fully annotated data to simulate interactions or non-interactively train models. *Online learning* methods require no annotated training data [19, 65, 85] or pre-train on a small fully-annotated dataset [6, 8]. This makes them particularly suitable for tasks where there is a limited amount of annotated data.

Recommendation. In cases where labels are scarce or costly to obtain, *application only* methods are unsuitable as they require fully annotated datasets for training. In contrast, *training only* and *online learning* methods are less dependent on this factor.

Q3. Does the task demand a specific model complexity?

The fundamental principles of the *training only* and *application only* methods do not mandate a specific model complexity. However, methods in the *online learning* node are limited to small models such as one-layer CNNs [65, 85] or a 2-layer U-Net [19] since the models are updated in real-time during application. We deem this disadvantage as "neutral" in Figure 32 as it constrains model architecture options without requiring high-end hardware like larger models.

Recommendation. If the task requires a complex model, avoid *online learning* methods as they rely on simpler models.

Q4. How diverse is the data during application?

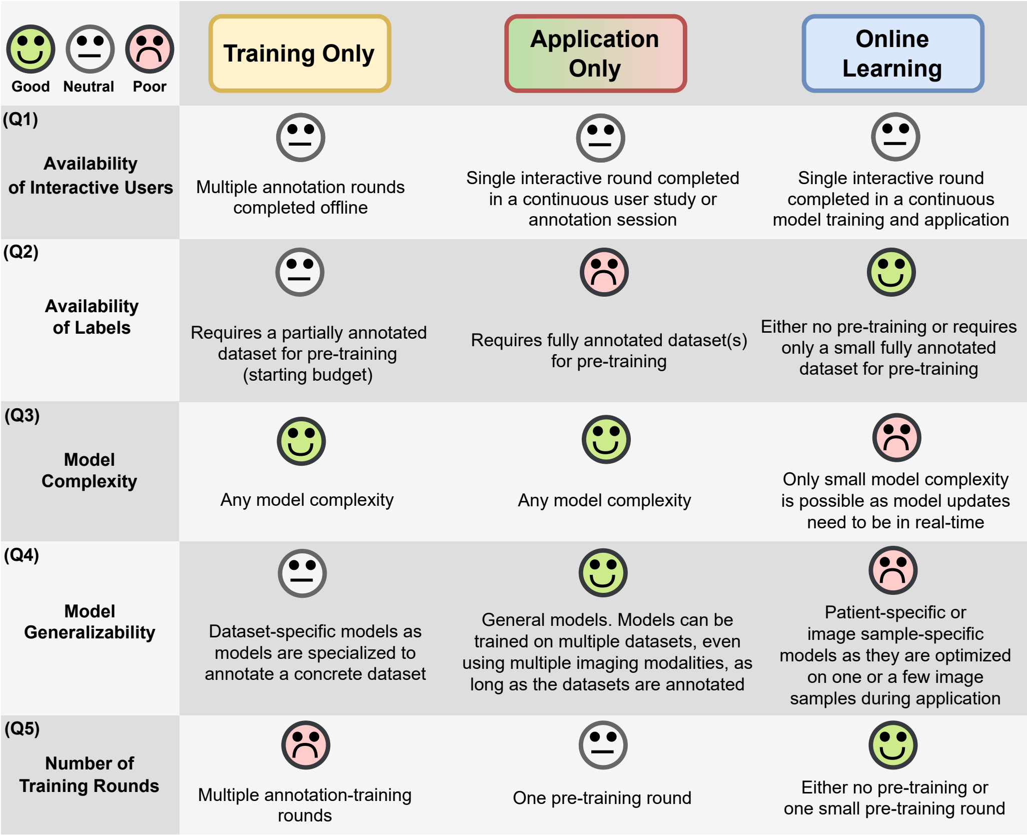
Methods in the *training only* paradigm are used in multiple annotation-training rounds to annotate one concrete dataset, tailoring them to that dataset [37, 62]. In contrast, *application only* methods are not limited to one dataset and may employ multiple training datasets, even from various imaging modalities [120]. The only requirement is that these datasets contain ground-truth labels, enabling either the simulation of interactions (green branch in Figure 31) or non-interactive pre-training (red branch in Figure 31). *Online learning* methods exhibit the most constrained generalization capability, as they are typically trained either on individual image samples [41, 65, 85] or only on samples obtained from a single patient [19].

Recommendation. For diverse application data, *application only* methods are most suitable as they can utilize training data from various sources or imaging modalities. *Training only* methods are suitable for partially annotated data aiming for full annotation, while *online learning* is ideal for single-patient or single-image-sample scenarios.

Q5. How time-critical is the deployment of the model?

The number of necessary training rounds places constraints on the hardware required for model training. *Training only* methods entail multiple annotation-training iterations, during which the model's predictions are iteratively refined and the model is trained multiple times until it reaches acceptable performance. This results in a slower deployment of the model for application and higher hardware demands. *Application only* methods require only a single pre-training round, and *online learning* methods either exclude pre-training altogether or require only a single small pre-training round. This makes them well-suited for scenarios with limited access to hardware.

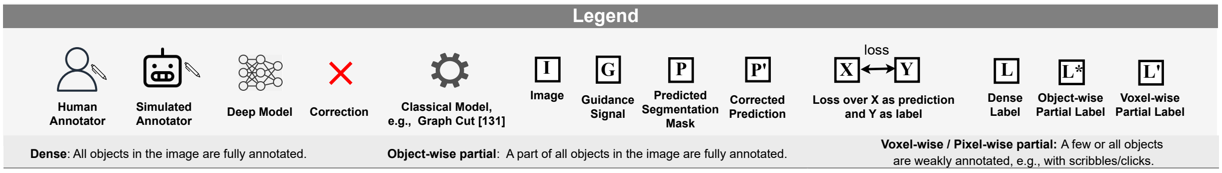
Recommendation. If the transition time from the training to the application phase needs to be short, we recommend designing a method following the *online learning* or *application only* paradigm as *training only* methods require multiple annotation-training rounds before their application.



🙂 Good 😐 Neutral ☹ Poor	Training Only	Application Only	Online Learning
(Q1) Availability of Interactive Users	😐 Multiple annotation rounds completed offline	😐 Single interactive round completed in a continuous user study or annotation session	😐 Single interactive round completed in a continuous model training and application
(Q2) Availability of Labels	😐 Requires a partially annotated dataset for pre-training (starting budget)	☹ Requires fully annotated dataset(s) for pre-training	🙂 Either no pre-training or requires only a small fully annotated dataset for pre-training
(Q3) Model Complexity	🙂 Any model complexity	🙂 Any model complexity	☹ Only small model complexity is possible as model updates need to be in real-time
(Q4) Model Generalizability	😐 Dataset-specific models as models are specialized to annotate a concrete dataset	🙂 General models. Models can be trained on multiple datasets, even using multiple imaging modalities, as long as the datasets are annotated	☹ Patient-specific or image sample-specific models as they are optimized on one or a few image samples during application
(Q5) Number of Training Rounds	☹ Multiple annotation-training rounds	😐 One pre-training round	🙂 Either no pre-training or one small pre-training round

Figure 32: Advantages and disadvantages of methods within the three main taxonomy nodes ***T1–T3*** regarding five specific demands.

After addressing *Q1-Q5*, users should gain a clearer understanding of which taxonomy node aligns with their specific use case. The next sections introduce the remaining questions *Q6-Q12*, offering guidance to navigate deeper into our taxonomy tree until reaching a leaf node.

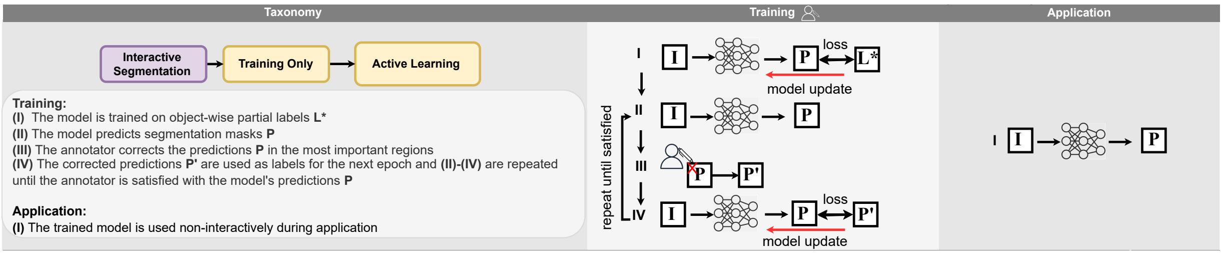


Dense: All objects in the image are fully annotated. **Object-wise partial**: A part of all objects in the image are fully annotated. **Voxel-wise / Pixel-wise partial:** A few or all objects are weakly annotated, e.g., with scribbles/clicks.

Figure 33: Visual legend illustrating the symbols and icons used throughout the taxonomy figures to indicate the training and application phases and human involvement.

6.1.3.2 *Visual Elements Overview*

In the next three sections, we discuss each taxonomy node in detail and provide visual examples of where human interactions occur, e.g., during training, application, or both (online learning). To simplify our visualizations of each node, we illustrate the training and application phases of the main taxonomy nodes, using icons to represent generic concepts, such as the input image. The diagram displays the involvement of a human annotator during either the training or application phase, or both (online learning). Figure 33 introduces all elements that are used in the visualizations in the next sections.



Training:
(I) The model is trained on object-wise partial labels **L***
(II) The model predicts segmentation masks **P**
(III) The annotator corrects the predictions **P** in the most important regions
(IV) The corrected predictions **P'** are used as labels for the next epoch and **(II)-(IV)** are repeated until the annotator is satisfied with the model's predictions **P**

Application:
(I) The trained model is used non-interactively during application

Figure 34: Illustration of the **T1.1. Active Learning** taxonomy node.

6.1.3.3 *Training Only*

The first taxonomy category ***T1. Training Only*** encompasses methods utilizing human interactions only during training, depicted as yellow nodes in our taxonomy tree in Figure 31. Reviewed methods within this paradigm all fall within one taxonomy node: ***T1.1. Active Learning***.

T1.1. Active learning models are first trained on a small labeled fraction of the dataset ("starting budget") and are subsequently applied to the unlabeled remainder of the dataset. Based on these predictions, the most informative samples for future training are identified, annotated, and added to the training data for the next iteration. This iterative training process continues until the annotator is content with the model's predictions. Afterward, the model may be used non-interactively on the application data without involving a human annotator, as seen in Figure 34.

6.1.3.4 *Application Only*

The second category of our taxonomy ***T2. Application Only*** encompasses models engaging with human annotators exclusively during the application stage, depicted as green and red nodes in Figure 31. During the training stage, these models either: 1) utilize simulated interactions generated by a simulated annotator ***(T2.1.)***, termed *robot user* in literature [122]; or 2) use no interactions ***(T2.2.)***. During the application stage, human users interact with these models by providing initial and/or iterative corrective interactions. To decide between the simulated and non-interactive methods, the user may consider the following question:

Q6. Are interactions simulated during training?

Simulated interactions (green nodes) facilitate the generation of interactions with predefined annotation behaviors, such as placing clicks at object centers or boundaries. This enhances the model's adaptation to specific behaviors so that it better leverages real user interactions during application. Non-interactive training (red nodes), on the other hand, does not integrate any prior knowledge about the annotation behavior. However, a drawback of simulated training, particularly when performed iteratively, is a longer training time due to the computational overhead from simulating interactions.

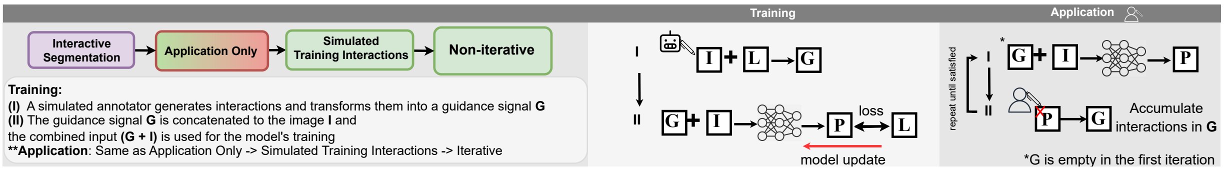
Recommendation. Simulated interactions specialize the model toward a specific interaction style, which could be helpful if, during application, annotators follow a specific annotation protocol. However, in scenarios where this is not important, or where a short training time is important, non-interactive training methods offer an alternative.

T2.1. Simulated training interactions circumvent the need for human annotators during the training process by simulating the annotation process using a *robot user.* This robot user mimics the behavior of a human annotator and relies on ground truth labels to simulate interactions only in the correct regions. In our taxonomy, we differentiate between non-iterative ***(T2.1.1.)*** and iterative simulation ***(T2.1.2.)*** and aid in selecting a method with the following question:

Q7. Is the model trained to refine its predictions?

Iterative simulations train models to refine their predictions with each new interaction, better aligning with real-world application scenarios where annotators continuously correct segmentations. However, non-iterative interaction simulations are computationally more efficient than iterative ones as they do not require multiple model predictions during training.

Recommendation. If the model should iteratively refine predictions and longer training is acceptable, iterative simulations are a suitable option. However, if the user interacts with the model only once or if the training time should be short, non-iterative methods are more favorable.



Training:
(I) A simulated annotator generates interactions and transforms them into a guidance signal **G**
(II) The guidance signal **G** is concatenated to the image **I** and the combined input **(G + I)** is used for the model's training
****Application**: Same as Application Only -> Simulated Training Interactions -> Iterative

Figure 35: Illustration of the **T2.1.1. Non-iterative simulation** taxonomy node.

T2.1.1. Non-iterative simulation methods generate all interactions at once in a single iteration and then transform them into a guidance signal, which is combined with the image. During training, there are no correction loops, whereas during the application stage, human annotators may iteratively correct the model's predictions, as illustrated in Figure 35. Non-iterative methods are further subdivided into the two subcategories *rule-based* ***(T2.1.1.1.)*** and *sampling-based* ***(T2.1.1.2.)***, depending on whether the interactions are generated through deterministic rules (e.g., the center of the largest connected component in the mask) or by randomly sampling the ground-truth mask, respectively. To decide between sampling- and rule-based approaches, we ask the following question:

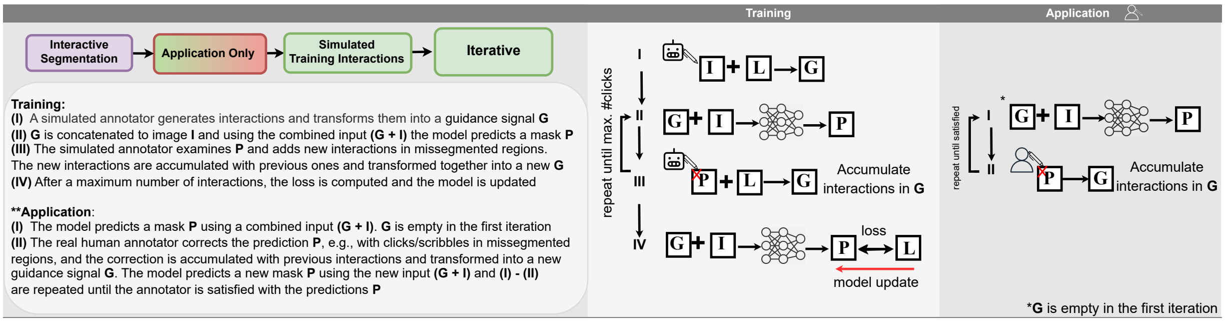
Q8. Is the simulation sampling- or rule-based?

In both iterative and non-iterative simulations, rule-based interaction generation provides precise control over annotation behavior, such as positioning clicks at the object's center. This protocol may be enforced during application, minimizing the user shift between training and application. On the other hand, sampling-based simulations introduce randomness in annotation, enhancing the model's capacity to generalize across diverse annotation styles and adapt to imperfect annotations.

Recommendation. Rule-based approaches are ideal when annotation styles are consistent and vary minimally, while sampling-based approaches excel when annotations are more flexible and interactions vary significantly among annotators.

The main difference between simulations is that ***T2.1.1.1. Non-iterative sampling-based*** methods sample the ground-truth labels to simulate interactions, whereas ***T2.1.1.2. Non-iterative rule-based*** methods employ deterministic rules to simulate interactions.

T2.1.2. Iterative simulation methods mimic the iterative nature of human interactions during application where the annotator repeatedly corrects the model prediction in a typical human-in-the-loop scenario, as seen in Figure 36. This loop is simulated by either sampling interactions from the missegmented regions or defining deterministic rules to choose each next interaction, e.g., choosing the center of the largest erroneous region. We refer to *Q8* for deciding between error sampling-based ***(T2.1.2.1.)*** and rule-based methods ***(T2.1.2.2.)***.


Training
Application
Interactive Segmentation
Application Only
Simulated Training Interactions
Iterative
repeat until max. #clicks
I
II
III
IV
I
L
G
P
Accumulate interactions in G
loss
model update
repeat until satisfied
*
I
II
Accumulate interactions in G
*G is empty in the first iteration

Training:
(I) A simulated annotator generates interactions and transforms them into a guidance signal **G**
(II) G is concatenated to image **I** and using the combined input **(G + I)** the model predicts a mask **P**
(III) The simulated annotator examines **P** and adds new interactions in missegmented regions.
The new interactions are accumulated with previous ones and transformed together into a new **G**
(IV) After a maximum number of interactions, the loss is computed and the model is updated

****Application**:
(I) The model predicts a mask **P** using a combined input **(G + I)**. **G** is empty in the first iteration
(II) The real human annotator corrects the prediction **P**, e.g., with clicks/scribbles in missegmented regions, and the correction is accumulated with previous interactions and transformed into a new guidance signal **G**. The model predicts a new mask **P** using the new input **(G + I)** and **(I) - (II)** are repeated until the annotator is satisfied with the predictions **P**

Figure 36: Illustration of the **T2.1.2. Iterative simulation** taxonomy node.

T2.1.2.1. Iterative error sampling-based methods sample interactions for the next iteration from missegmented regions. We distinguish between uniform ***(T2.1.2.1.1.)*** and distance transform-based ***(T2.1.2.2.)*** iterative sampling in our taxonomy by asking the following:

Q9. How are interactions sampled?

Unlike non-iterative simulations, iterative methods involve two sampling types: 1) uniform; 2) and distance transform-based sampling. ***T2.1.2.1.1. Iterative uniform sampling*** approaches sample new interactions with an equal probability of landing in any of the missegmented pixels/voxels. ***T2.1.2.1.2. Iterative distance transform-based sampling*** methods apply a distance transform over the missegmented regions, generating a distance map that serves as a sampling distribution for new interactions. As a result, these approaches prioritize sampling new interactions primarily in the central regions of the connected components of the errors.

Recommendation. Distance transform-based sampling is preferable when interactions are expected in central regions, while uniform sampling is suited for interactions anywhere in the label.

T2.1.2.2. Iterative rule-based approaches utilize a deterministic rule to generate an interaction at each iteration. We differentiate four types of rules: ***T2.1.2.2.1. Center of largest error*** methods which use the center of the largest error region as the next click with the assumption that it is the most intuitive choice; ***T2.1.2.2.2. Error skeletonization*** which simulates iterative scribbles as an alternative to iterative clicks. Similar to the central error clicks, error skeletonization generates scribbles that are positioned in the central regions of the object; ***T2.1.2.2.3. Multiple custom rules*** which are methods that apply multiple custom rules that are specific to the application; ***T2.1.2.2.4. Worst vertex/slice correction*** involves identifying and selecting the worst vertex or slice at each iteration and incorporating its ground-truth value as a guidance signal. Hence, we ask the following to select a node:

Q10. What rules are used?

There are no clear advantages as each node is tailored to a specific interaction type, e.g., center of largest error for clicks, error skeletonization for scribbles, worst vertex correction for polygons, and multiple custom rules for methods using multiple interaction types.

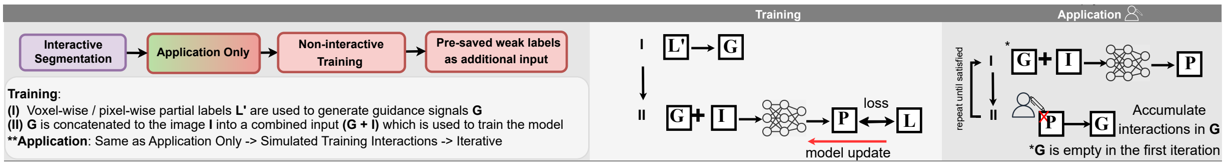
Recommendation. We recommend selecting rule-based nodes depending on the interaction type, e.g., clicks, scribbles, bounding box, polygon vertices.

T2.2. Non-interactive training methods opt to exclude interactions during the training stage and, instead, adopt a standard non-interactive training approach. These methods are marked in red in Figure 31. Based on their approach, these methods either incorporate additional weak labels during training ***(T2.2.1.)*** or post-process the model prediction during the application ***(T2.2.2.)***. To select between these two types of methods, we ask the following:

Q11. Does the training data have additional weak labels?

Pre-saved weak labels provide similar advantages as non-iterative simulations, enabling the model to adapt to specific annotation behaviors by incorporating weak labels as additional model inputs. However, they require manual annotations, which is costly for large datasets, even for weak labels. In contrast, post-processing during application does not require any annotation efforts but does not incorporate prior knowledge about the annotation style into the model.

Recommendation. Pre-saved weak label methods are suitable when weak labels exist in the training data, aiding in adapting the model to specific annotation behaviors. Alternatively, if weak labels are absent or annotation behavior is not crucial, post-processing provides an efficient solution.

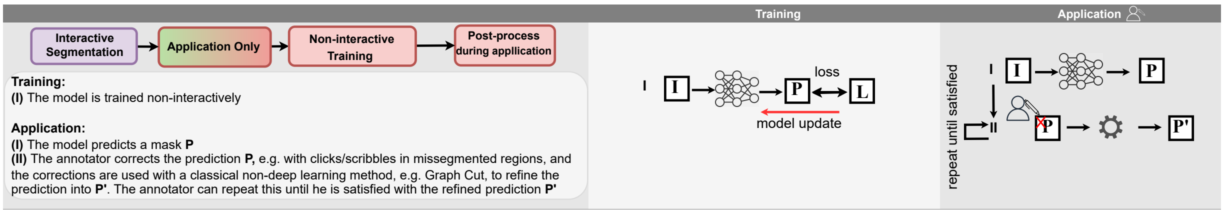


Training:
(I) Voxel-wise / pixel-wise partial labels **L'** are used to generate guidance signals **G**
(II) G is concatenated to the image **I** into a combined input **(G + I)** which is used to train the model
****Application**: Same as Application Only -> Simulated Training Interactions -> Iterative

Figure 37: Illustration of the **T2.2.1. Pre-saved weak labels as additional input** taxonomy node.

T2.2.1. Pre-saved weak labels as additional input methods utilize additional weak labels during training. However, instead of using the weak labels as supervisory signals as done in weakly supervised learning [129], the weak labels are transformed into guidance signals as seen in Figure 37.

T2.2.2. Post-processing during application methods adopt non-interactive training and integrate post-processing techniques to combine model predictions with user interactions during application as seen in Figure 38.



Training:
(I) The model is trained non-interactively

Application:
(I) The model predicts a mask **P**
(II) The annotator corrects the prediction **P,** e.g. with clicks/scribbles in missegmented regions, and the corrections are used with a classical non-deep learning method, e.g. Graph Cut, to refine the prediction into **P'**. The annotator can repeat this until he is satisfied with the refined prediction **P'**

Figure 38: Illustration of the **T2.2.2. Post-processing during application** taxonomy node.

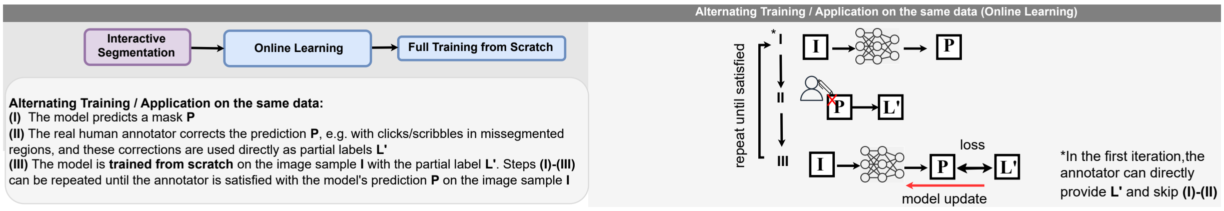
6.1.3.5 *Online Learning*

The third category in the taxonomy tree ***T3. Online Learning*** encompasses methods that undergo real-time training or fine-tuning directly on the data they are finally applied to. Methods in this paradigm produce on-the-fly predictions and allow annotators to make immediate corrections with minimal or no latency between corrections, model updates, and new predictions. In our taxonomy, we differentiate between full training ***(T3.1.)*** and fine-tuning ***(T3.2.)*** based on the number of updated model parameters and we ask the following to help select a method:

Q12. What part of the model is trained?

In *online learning*, full model training relies solely on interactions during application as labels, eliminating the need for any manual annotations for pre-training. However, fine-tuning necessitates a small pre-training dataset to initialize the model. Despite this, full training approaches start with entirely random initial predictions, requiring more interactions to achieve performance levels comparable to fine-tuning approaches.

Recommendation. Full model training is suitable in cases where there is no labeled training data and the model is directly trained on the application data, whereas fine-tuning is the better option if a small labeled dataset is already present.


Alternating Training / Application on the same data (Online Learning)
Interactive Segmentation
Online Learning
Full Training from Scratch
repeat until satisfied
*
I
II
III
loss
model update

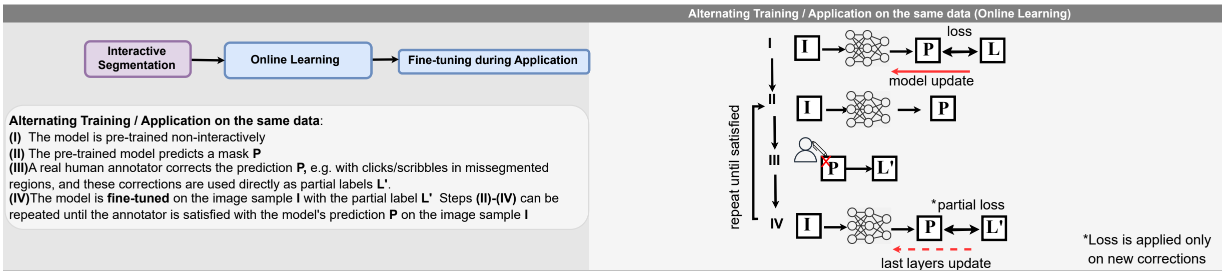
Alternating Training / Application on the same data:
(I) The model predicts a mask **P**
(II) The real human annotator corrects the prediction **P**, e.g. with clicks/scribbles in missegmented regions, and these corrections are used directly as partial labels **L'**
(III) The model is **trained from scratch** on the image sample **I** with the partial label **L'**. Steps **(I)-(III)** can be repeated until the annotator is satisfied with the model's prediction **P** on the image sample **I**

*In the first iteration,the annotator can directly provide **L'** and skip **(I)-(II)**

Figure 39: Illustration of the **T3.1. Full training from scratch** taxonomy node.

T3.1. Full training from scratch models do not use any pre-training and are trained entirely on the data on which they are finally applied, as seen in Figure 39. These models use the user interactions as the only labels, update their parameters in real-time, and predict again so that the user may correct them again until they are satisfied with the prediction's quality.

T3.2. Fine-tuning during application online learning models utilize a pre-trained model and only fine-tune it on the application data using human interactions, as seen in Figure 40.


Alternating Training / Application on the same data (Online Learning)
Interactive Segmentation
Online Learning
Fine-tuning during Application
repeat until satisfied
I
II
III
IV
I
P
L
L'
loss
model update
*partial loss
last layers update
*Loss is applied only on new corrections

Alternating Training / Application on the same data:
(I) The model is pre-trained non-interactively
(II) The pre-trained model predicts a mask **P**
(III)A real human annotator corrects the prediction **P,** e.g. with clicks/scribbles in missegmented regions, and these corrections are used directly as partial labels **L'**.
(IV) The model is **fine-tuned** on the image sample **I** with the partial label **L'** Steps **(II)-(IV)** can be repeated until the annotator is satisfied with the model's prediction **P** on the image sample **I**

Figure 40: Illustration of the **T3.2. Fine-tuning during application** taxonomy node.

6.2 REVIEW FINDINGS

In this section, we present our findings on the prevalent trends observed during the review of the 121 reviewed papers. We delve into the implications of these trends and the potential factors contributing to them. Through this analysis, we aim to provide a comprehensive depiction of the current landscape within the medical interactive segmentation domain.

6.2.1 *Current Trends among Reviewed Studies*

6.2.1.1 *Segmentation Targets*

Segmentation targets fall into two main categories: (1) anatomical structures and cells, and (2) pathologies. This distinction depends on whether a method primarily focuses on anatomy, pathology, or both (noted in 7 of 121 studies). Figure 41 summarizes the distribution across targets. Common anatomical regions include the brain, prostate, and cardiac structures, along with abdominal organs such as the liver, spleen, kidney, pancreas, stomach, and gallbladder. Less frequent targets comprise thoracic organs (lungs, aorta, esophagus), whole-body structures (bones, vessels), and regions grouped under “Other,” such as lymph nodes, the Z-line, spine, cartilage, and skin. Microscopy and OCT-based techniques typically address cell segmentation, focusing on blood cells, testicular cells, neurons, or nuclei.

Pathology-focused targets are less diverse, concentrating mainly on brain (n=21) and liver (n=12) lesions, reflecting the influence of public datasets like *BraTS* [162], *MSD* [163], and *LiTS* [164]. Beyond these, COVID-19 lung lesions (X-rays), skin lesions (dermoscopy), and colon polyps (colonoscopy) are common. Other pathologies include lung, breast, kidney, and thyroid cancers, with rarer cases (in “Other”) covering head and neck, cervical, pancreatic, prostatic, and esophageal cancers, as well as hematomas and foot ulcers.

6.2.1.2 *Imaging Modalities*

Radiological imaging dominates, with CT (n=65) and MRI (n=42) featuring most prominently, largely due to public datasets from challenges like *MSD* [163] and *BraTS* [162]. Ultrasound

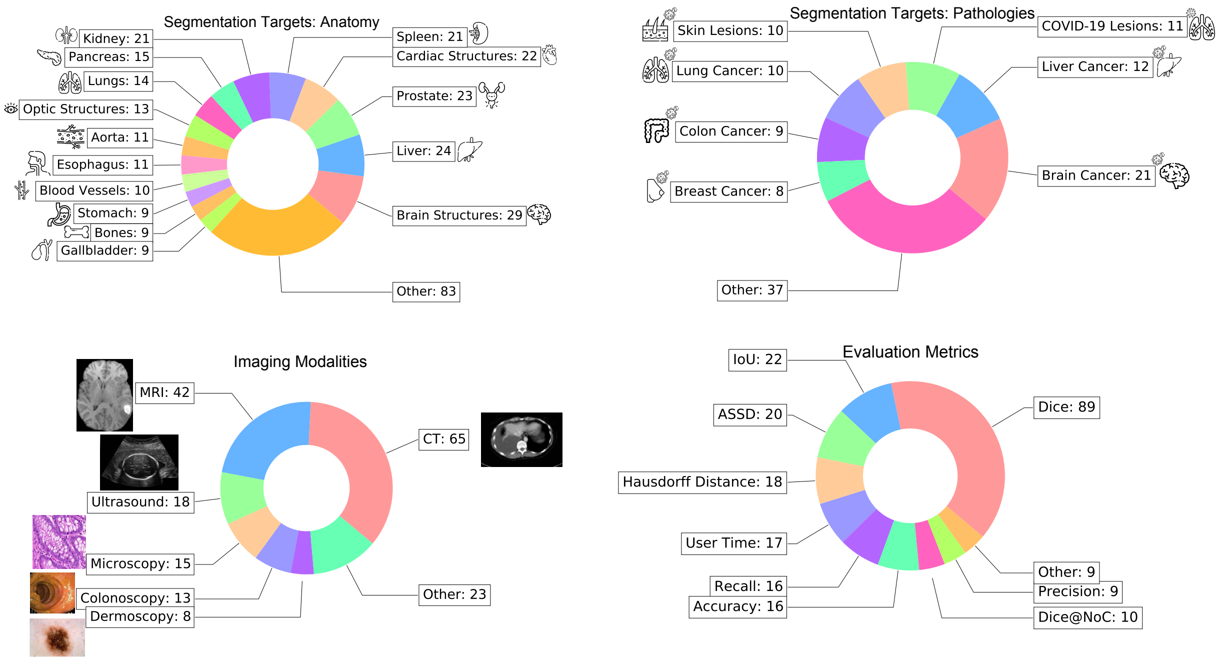


Figure 41: Distribution of segmentation targets in the anatomy (top left) and in the pathology (top right), imaging modalities (bottom left), and evaluation metrics (bottom right) among all reviewed papers. The numbers represent the number of papers in that category. The icons on the top row are designed by Flaticon.com.

appears in 18 studies, mainly for cardiac, mammographic, or fetal imaging. Microscopy (n=15) is used primarily for tumor or cancer cell segmentation. Colonoscopy is dedicated to colon cancer and polyp segmentation, while dermoscopy focuses on skin lesions. Less common modalities include OCT, X-ray, fundus imaging, and PET/CT.

6.2.1.3 *Evaluation Metrics*

The choice of evaluation metrics critically determines the reliability and interpretability of segmentation results. Recent analyses highlight major pitfalls in metric usage [170], particularly regarding the Dice Similarity Coefficient (DSC). The "Metrics Reloaded" framework [165] addresses these issues by recommending multi-metric evaluation to capture diverse failure modes. Figure 41 summarizes the metrics used across reviewed studies. As expected, DSC is most common (n=89), yet 29 studies rely on it exclusively, and 19 redundantly report both DSC and Intersection over Union (IoU). Despite its clinical relevance and endorsement by "Metrics Reloaded," the Normalized Surface Distance [171] appears in only 2 of 121 studies.

User-centered metrics remain underrepresented despite their importance for human-in-the-loop methods. Reported examples include User Time (n=17), measuring annotation duration, and Dice@NoC (n=10), assessing accuracy after a fixed number of interactions. The

"Other" category includes usability measures such as NASA-TLX [147] and the System Usability Scale [148], though these are rarely applied.

6.2.1.4 *Emergence of Foundation Models*

The Segment Anything Model (SAM) [137], introduced in early 2023, marked a shift toward foundation models in segmentation through large-scale training on over one billion masks. Although trained primarily on 2D natural images (SA-1B), SAM has been effectively adapted to medical domains across 2D (e.g., dermoscopy, fundus) and 3D modalities (e.g., CT, MRI, PET/CT) via targeted fine-tuning [120]. For 3D data, adaptations often employ 2D axial slices or specialized 2D-to-3D adapters [118]. SAM's strong generalization, particularly for 2D modalities, and its efficient bounding-box prompting have spurred rapid progress in interactive medical segmentation. Within months of its release, 29 medical SAM adaptations were published, underscoring its impact. Leveraging SAM's zero-shot and cross-domain capabilities, several approaches now evaluate performance across 30 or more public medical datasets [119, 120], highlighting the growing influence of foundation models in the field.

6.2.1.5 *Reproducibility and Availability*

An increasing number of studies release code, replication instructions, and occasionally pretrained weights, fostering transparency and reuse. Open-source frameworks such as *MONAI Label* [149], *AnatomySketch* [70], *RIL-Contour* [150], *BioMedisa* [151], *MITK* [173], and *PyMIC* [174] further support development and deployment. Non–deep learning tools like *ilastik* [168], *ITK-Snap* [169], and the method by Li et al. [172] remain widely used. This trend is reinforced by the availability of open challenge datasets from platforms such as *Kaggle* (kaggle.com), *Grand Challenge* (grand-challenge.org), and *Synapse* (synapse.org), collectively promoting collaboration and advancing the field.

6.2.2 *How Existing Approaches Compare to Each Other*

We examined how current methods benchmark against existing approaches, as scientific progress relies on evaluating gains over established baselines. Figure 42 summarizes comparison practices across reviewed studies. Notably, 46 of 121 studies *do not compare against any prior work*. Others evaluate only against classical (pre-2016, non–deep learning) methods (n=6) or solely against *DeepIGeoS* [5] (n=3). A further 37 studies compare exclusively to non-medical interactive models such as *DIOS* [140], *Polygon-RNN* [175], *DEXTR* [176], *Latent Diversity* [179], *BRS* [178], *f-BRS* [177], and SAM [137]. Even methods belonging to the same taxonomy node (Figure 42) rarely compare with one another. Moreover, the rapid adoption of SAM has further limited cross-comparison, as most SAM-based studies evaluate only against the original non-medical SAM. This widespread lack of systematic benchmarking contrasts sharply with the positive reproducibility trends shown in Figure 29.

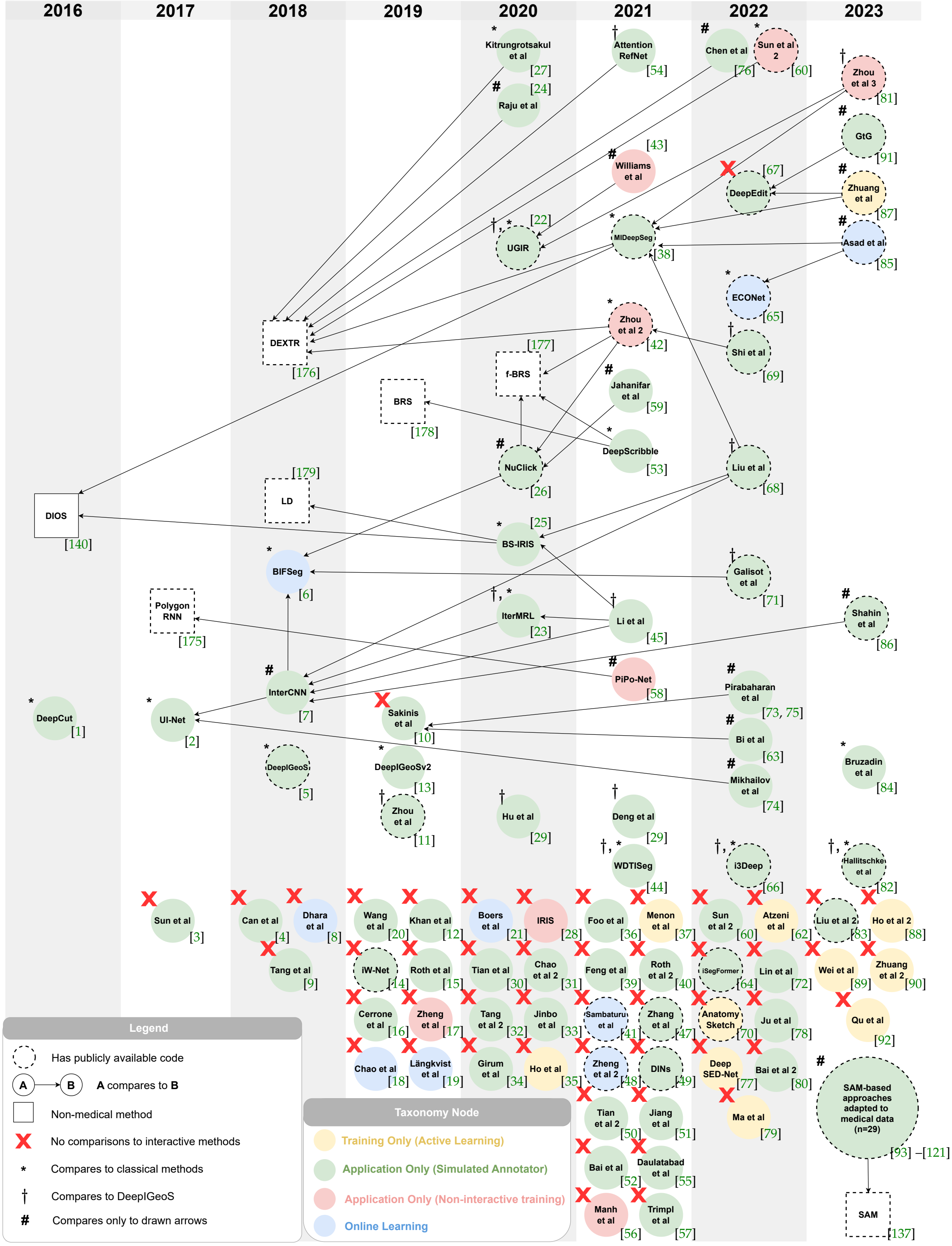


Figure 42: Comparison graph of all the reviewed methods. Nodes are ordered by initial submission year left to right. Classical methods denote non-deep learning-based interactive methods before 2016. The star (∗) and the dagger (†) are introduced to reduce the visual load in the figure caused by too many arrows.

6.3 OPPORTUNITIES TO IMPROVE THE FIELD

Our review highlights a pivotal challenge in the field: a discernible deficiency in scientific rigor in method evaluation. This challenge is evident in various aspects that we discuss in the following, alongside opportunities to address them.

6.3.1 *Missing Baselines and Scattered Comparisons*

The absence of consistent baselines and scattered comparisons across studies is a major issue. Frequently, new methods are not compared with previous work, possibly due to a lack of awareness of other methods or no established evaluation protocols for interactive segmentation.

Opportunities: First, we hope that our taxonomy tree functions as a navigational tool, aiding researchers in categorizing their approaches and guiding them towards relevant existing methods. Second, the emergence of generalizing models like SAM [137] is a promising trend towards foundational baselines that allow for out-of-the-box comparisons under a uniform protocol. This approach can shift the field towards more structured and systematic improvements, similar to the effects of *nnU-Net* [136] in the realm of non-interactive medical segmentation, which, due to its out-of-the-box functionality, serves as a strong and standardized baseline in the field [146].

6.3.2 *No Standardized Benchmarking Datasets*

The lack of established benchmarking protocols across datasets and tasks in interactive medical image segmentation is a significant barrier. This gap impedes the objective evaluation and comparison of interactive models, which results in an inconsistent literature landscape with no definite state of the art. A major challenge in benchmarks for the medical domain is that simulating interactions requires expert knowledge, as annotators develop their styles over years of experience with specific tasks. Therefore, creating meaningful benchmarks involves simulating interactions that reflect this expertise.

Opportunities: The domain of non-medical interactive segmentation, particularly with natural images, has addressed this issue by leveraging extensively validated benchmark datasets like *GrabCut* [141], *DAVIS* [142], *Pascal VOC* [143], *SBD* [144], and *Berkeley* [145]. Moreover, these datasets are coupled with well-defined evaluation protocols and metrics, streamlining fair and systematic comparisons with previous research. A potential remedy for the fragmented nature of comparisons within the medical interactive segmentation field entails the establishment of a curated selection of the most exemplary datasets tailored to specific tasks and imaging modalities, complete with well-defined evaluation protocols. When designing a medical interactive benchmark, we suggest evaluating models with various annotation styles to ensure effective use by different annotators. For natural images, benchmarks typically use the "center click in the largest error" protocol to simulate clicks [177], [179]. However, for medical tasks, it is important to include multiple diverse simulated annotators in each benchmark to test whether interactive models perform reliably when used by annotators focusing on different types of errors. Such an approach would furnish researchers with a systematic framework for assessing their methodologies and documenting enhancements over prior methods.

6.3.3 *Lack of Adequate and Standardized Evaluation Metrics*

In the current landscape of deep interactive medical image segmentation, there are two significant challenges related to metric selection. The first prevalent issue is the over-reliance on a single metric for evaluating segmentation performance. As pointed out in [165, 170], this approach is too narrow and often fails to adequately capture the complexity and nuances of segmentation accuracy. Second, there is a conspicuous absence of user-centric metrics in evaluations. These metrics are essential to understand how effectively an interactive segmentation tool meets the practical needs and scenarios of its users, especially in the medical imaging context.

Opportunities: By adopting the comprehensive guidelines of "Metrics Reloaded" for metric selection, researchers can ensure a more holistic evaluation of segmentation methods. This would involve using a diverse set of metrics that together provide a more complete picture of a method's performance. In addition to technical metrics, emphasizing user-centric metrics in evaluations is crucial. These metrics focus on: (1) annotation efficiency, measuring how quickly an image is annotated (e.g., user time per image or #Inter@90); and (2) annotation efficacy, measuring how well interactions are utilized (e.g., Dice Score after 10 clicks, DSC@10, or Consistent Improvement - the percentage of interactions that lead to improved segmentation). We believe both categories should be included in every benchmark to comprehensively evaluate interactive models.

This focus will shed light on the usability and practical effectiveness of interactive segmentation methods from the perspective of end-users, which is particularly important in clinical applications. Our review found that iterative interactive methods are the most popular taxonomy category (n=61). When designing benchmarks, we suggest evaluating models with multiple robot users to account for the variability in annotation styles as discussed in Section 6.3.2 and to report the performance before the first interaction to assess how the task is addressed with non-interactive methods.

6.4 CHAPTER CONCLUSION

Our systematic review and taxonomy serve as a key resource for researchers and practitioners in deep interactive medical image segmentation. Researchers can more easily identify relevant studies to strengthen their methodologies, while practitioners can select approaches suited to their specific tasks. Beyond mapping trends, our review exposes major challenges—most notably, a lack of scientific rigor and standardized evaluation. Addressing these issues through systematic benchmarking is essential for advancing reliable and effective medical interactive segmentation methods.

We summarize the scientific impact of this chapter in four key contributions:

Contribution 1: A comprehensive taxonomy for deep interactive medical segmentation allowing users to quickly comprehend existing approaches (*TPAMI 2024* [209]).

Contribution 2: A question-based recommendation system *Q1-Q12* to select the best-fitting interactive method for a specific task (*TPAMI 2024* [209]).

Contribution 3: Identification of key challenges in interactive segmentation, including the lack of standardized benchmarks, datasets, and baselines required for fair and consistent method comparison (*TPAMI 2024* [209]).

Contribution 4: Establishment of theoretical foundations through the definition of core concepts: *guidance signal*, *robot user*, and the distinction between training and application phases (*TPAMI 2024* [209]). This contribution is also discussed in greater detail in Chapters 3 and 5.

Our review unveils the fragmented state of the medical interactive segmentation field, but also discusses opportunities to bring it forward by focusing on standardizing comparisons between approaches through established benchmarks. As this is an entire research field, standardization should arise from the community. Hence, we ask the following:

How can we form an interactive segmentation community?

In the next chapter, we explore building an interactive segmentation community by co-organizing global challenges with research centers. This initiative fosters collaboration, transparency, and reproducibility, while establishing shared benchmarks, best practices, and a foundation for collective progress in the field.

7 THE INTERACTIVE SEGMENTATION INITIATIVE

In the previous chapter, we identified critical gaps in the field of medical interactive segmentation, most prominently the absence of standardized benchmarks that enable fair and reproducible comparison between methods. We view this as a global issue that can only be addressed through coordinated, community-driven collaboration rather than isolated individual efforts. In this chapter, we present the concrete steps we have taken toward this goal by establishing the *Interactive Segmentation Initiative* - a collaborative effort uniting research groups worldwide through the joint organization of international interactive segmentation challenges. Section 7.1 defines the concept of a *medical segmentation challenge* and explains its structure and organization. Section 7.2 discusses the formation of the *Interactive Segmentation Initiative*, which expands established non-interactive segmentation challenges into interactive ones. Section 7.3 describes how we unified these challenges through shared evaluation metrics, interaction simulations, and baseline models to enable meaningful cross-challenge comparisons. Finally, Section 7.4 presents the insights we gained from these consolidated results and highlights the key components for a successful interactive segmentation model.

This **entire chapter** is based on our co-organization of the segmentation challenges: (1) Automated Lesion Segmentation in Whole-Body PET/CT (autoPET)[1]; (2) ToothFairy3[2] as part of the Oral and Dental Image aNalysis challenges (ODIN)[3]; (3) Triphasic-aided Liver Lesion Segmentation (TriALS)[4]; and (4) CVPR 2025: Foundation Models for Interactive 3D Biomedical Image Segmentation (MedSegFM)[5].

7.1 WHAT IS A MEDICAL SEGMENTATION CHALLENGE?

Medical competitions, commonly referred to in the medical image analysis community as *challenges*[6], are community-driven initiatives designed to advance research through standardized,

[1] https://autopet-iv.grand-challenge.org/
[2] https://toothfairy3.grand-challenge.org/
[3] https://odin-workshops.org/2025/challenges.html
[4] https://www.synapse.org/Synapse:syn65878273/wiki/631556
[5] https://www.codabench.org/competitions/5263/
[6] https://miccai.org/index.php/special-interest-groups/challenges/

transparent, and reproducible evaluation. While challenges may address a variety of tasks, including classification, registration, or detection, this thesis focuses on segmentation challenges.

The key idea behind a challenge is to crowdsource the development of solutions by inviting the research community to compete on a shared task, rather than solving it in isolation. To ensure fairness, submissions are evaluated through centralized platforms such as *Grand Challenge*[7], *Codabench* [290], or *Synapse*[8], which provide identical evaluation conditions for all participants and maintain consistent, unbiased assessment [246].

A typical challenge proceeds in two main phases: *model development* and *final model testing*, as illustrated in Figure 43. During the development phase, participants receive a training dataset with ground-truth annotations and iteratively refine their models. Submissions are evaluated on a hidden validation set via an automated server, which provides feedback and updates a live leaderboard to display provisional rankings. In the final testing phase, participants submit their best-performing models for evaluation on a fully hidden test set. Once all submissions are scored, the final rankings are published, and results are presented during dedicated challenge workshop sessions at conferences. We define a *medical segmentation challenge* in Definition 3:

Definition 3: A *medical segmentation challenge* is a structured, community-driven benchmarking competition in which participants develop and evaluate segmentation algorithms on shared datasets under standardized conditions. It typically comprises two phases: (1) a *model development phase*, where participants train and validate their models on provided data, and (2) a *final testing phase*, where the best models are evaluated on a hidden test set to determine the final ranking under fair, transparent, and reproducible conditions.

Beyond as competitions, challenges serve as catalysts for collaboration and knowledge exchange. Their structure encourages innovation [146] while fostering enduring research communities. Organizers and participants often co-author post-challenge publications, and accompanying workshops provide platforms for open discussion and dissemination of best practices.

Over time, medical segmentation challenges have proven to be powerful engines for community formation and methodological standardization. Their open, collaborative structure not only stimulates innovation but also naturally drives the emergence of shared evaluation metrics, datasets, and best practices across research areas. These collective outcomes have established challenges as the de facto mechanism for setting standards within the broader medical image analysis community. Motivated by this success, we extend the challenge paradigm to the domain of interactive segmentation - an area still fragmented by the absence of unified benchmarks and evaluation protocols, as we saw in Chapter 6. By introducing interactive segmentation challenges, we aim to translate the proven strengths of the challenge format—standardization, transparency, and global collaboration—into a framework that can systematically address the field's most pressing issues and accelerate scientific progress.

[7] https://grand-challenge.org/

[8] https://www.synapse.org/

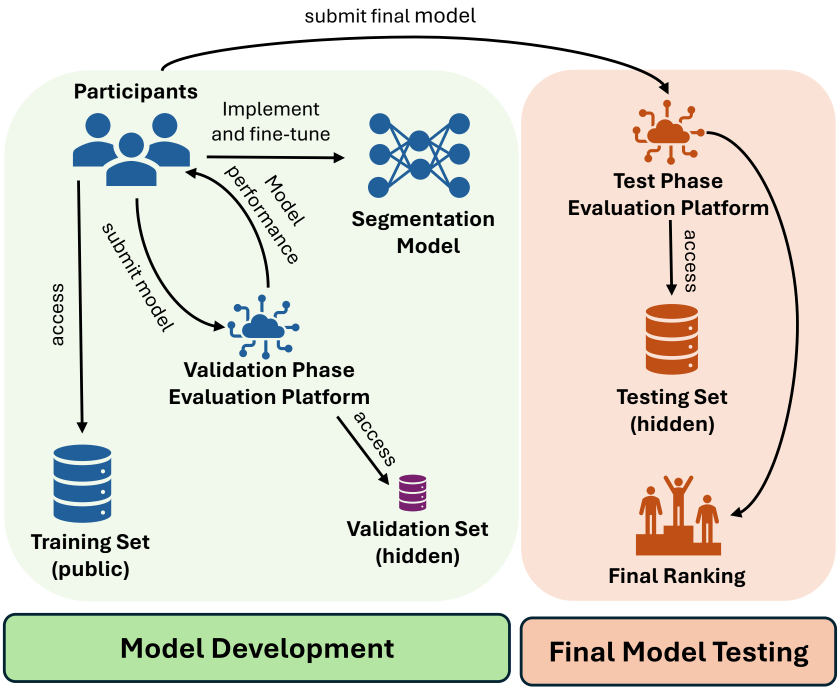


Figure 43: An example of a challenge's two phases - model development and final model testing.

7.2 FORMING AN INTERACTIVE SEGMENTATION COMMUNITY

To establish a unified interactive segmentation community, we strategically collaborated with long-standing medical image analysis challenges that demonstrated both continuity and relevance for future iterations. We prioritized challenges that remained unsolved, where previous winning methods achieved suboptimal segmentation performance, indicating room for improvement through interactive approaches. Furthermore, we focused on challenges hosted at leading venues such as *MICCAI* and *CVPR* to ensure both visibility and broad community engagement.

To make our initiative broadly applicable and easily adoptable, we designed a general framework, illustrated in Figure 44, that seamlessly integrates with existing challenge infrastructures. We distinguish between components already established within the challenges: (1) datasets; (2) dataset splits: (3) evaluation platforms (e.g., *Codabench* [290]); and (4) underlying tasks (e.g., whole-body lesion segmentation in *autoPET* [154]), and the extensions introduced through our initiative. Our contribution adds an interactive segmentation layer *without* altering the main challenge task, imposing additional implementation burden, or introducing legal or financial complications for organizers.

As depicted in Figure 44, our initiative focuses on four key extensions: (1) **interaction simulation**, providing synthetic user interactions to augment existing datasets; (2) **interactive**

evaluation metrics, transforming established metrics into their interactive counterparts while preserving comparability; (3) **interactive leaderboard**, containing results across multiple challenges to identify trends and highlight characteristics of strong interactive models; and (4) **a shared baseline**, implemented through our *SW-FastEdit* model from Chapter 4 adapted to each challenge, serving as a reference for all participants. This coordinated approach enables the creation of a cohesive ecosystem in which interactive segmentation methods can be developed, evaluated, and compared consistently across diverse tasks.

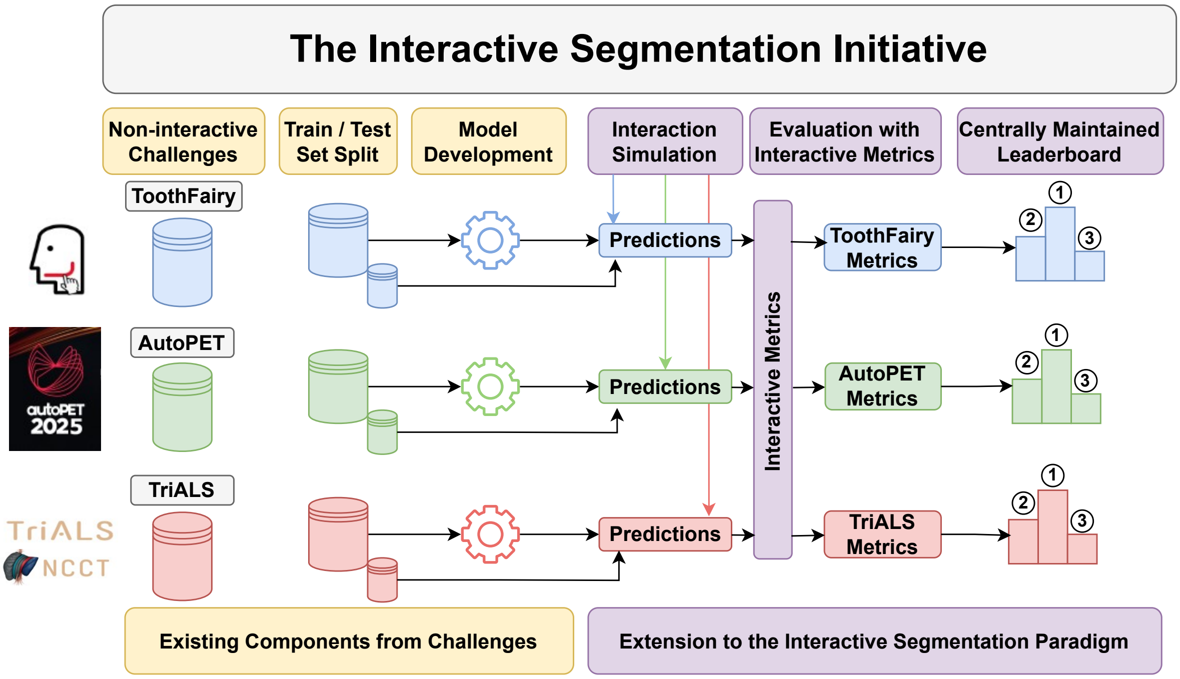

The Interactive Segmentation Initiative
Non-interactive Challenges
Train / Test Set Split
Model Development
Interaction Simulation
Evaluation with Interactive Metrics
Centrally Maintained Leaderboard
ToothFairy
AutoPET
TriALS
autoPET 2025
TriALS
NCCT
Predictions
Predictions
Predictions
Interactive Metrics
ToothFairy Metrics
AutoPET Metrics
TriALS Metrics
1
2
3
Existing Components from Challenges
Extension to the Interactive Segmentation Paradigm

Figure 44: Our extension of non-interactive challenges to the interactive segmentation paradigm in our initiative *without* altering the main objectives and structures of the challenges.

After getting in touch with existing challenges, we received positive feedback and willingness to collaborate from four challenges listed in Table 17 and illustrated in Figure 45. This leads to a global unified effort spanning 4 continents and over 13 research groups, bringing the community together to solve the research gap in interactive segmentation.

7.3 CONSOLIDATING DIFFERENT SEGMENTATION TASKS

7.3.1 *On the Diversity of Segmentation Challenges*

The four challenges we co-organized span a diverse range of imaging modalities and clinical applications: dental imaging (ToothFairy3), whole-body PET/CT scans (autoPET), portal-

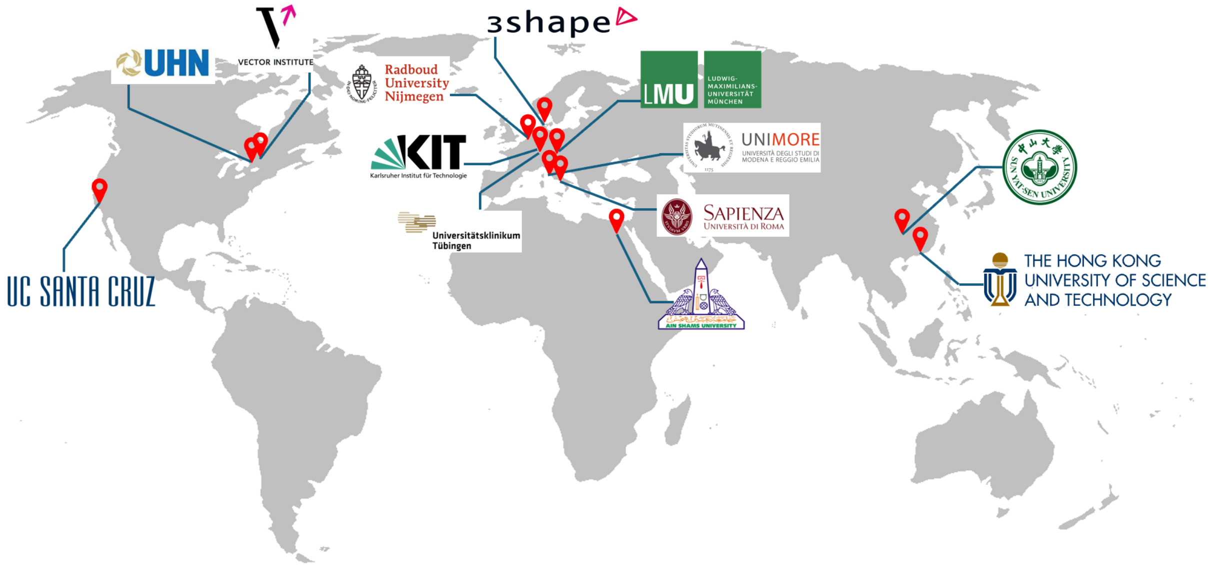

UHN
VECTOR INSTITUTE
3shape
Radboud University Nijmegen
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
KIT Karlsruher Institut für Technologie
UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA
SUN YAT-SEN UNIVERSITY
Universitätsklinikum Tübingen
SAPIENZA UNIVERSITÀ DI ROMA
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY
UC SANTA CRUZ
AIN SHAMS UNIVERSITY

Figure 45: Involved research groups in the organization of the four segmentation challenges.

Challenge	Image Modality	Target(s)	Conference	Organizing Research Groups
autoPET	PET/CT	Whole-body lesions	MICCAI 2025	University Hospital Tübingen, Germany; LMU University Hospital, Germany
ToothFairy3	Cone-Beam CT	Inferior Alveolar Canals	MICCAI 2025	University of Modena and Reggio Emilia, Italy; Radboud University, Netherlands; Sapienza University of Rome, Italy; Affidea, Italy; 3Shape A/S, Denmark
TriALS	Portal-venous CT	Liver Lesions	MICCAI 2025	The Hong Kong University of Science and Technology, Hong Kong SAR; Sun Yat-Sen University, Guangzhou, China; Ain Shams University, Cairo, Egypt
MedSegFM	CT, MRI, PET/CT, 3D Ultrasound, 3D Microscopy	Many (> 100)	CVPR 2025	University Health Network, Vector Institute, Toronto, Canada; UC Santa Cruz, USA

Table 17: Overview of the challenges co-organized as part of the Interactive Segmentation Initiative. We represent the Karlsruhe Institute of Technology (KIT), Germany, as co-organizers for all four challenges.

venous CT for liver lesions (TriALS), and 3D multimodal biomedical imaging for interactive foundation models (MedSegFM). Here, we briefly summarize the focus and objectives of each challenge.

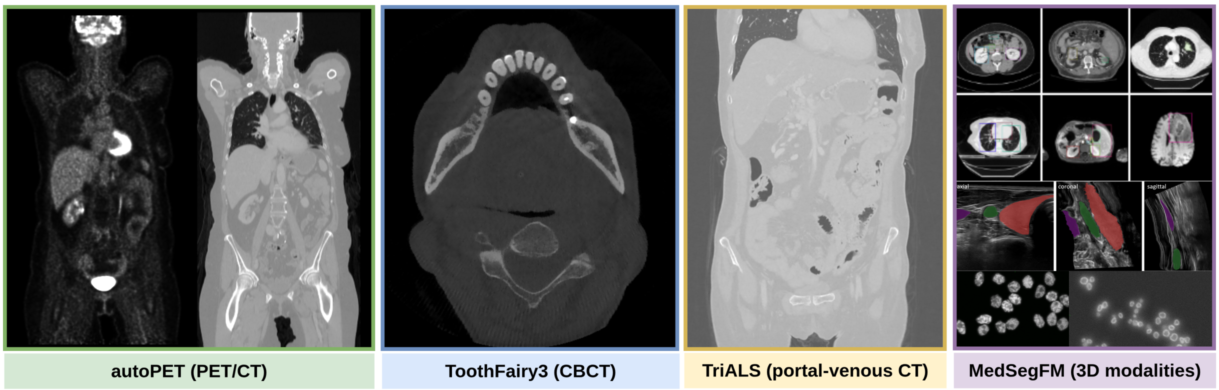
autoPET addresses automated tumor lesion segmentation in whole-body PET/CT imaging, aiming to advance quantitative analysis for oncologic diagnosis and therapy assessment [154].

The dataset includes paired PET and CT scans, primarily acquired using the FDG tracer, which reflects glucose metabolism and highlights tumor activity, as well as PSMA PET/CT scans that visualize prostate cancer lesions through prostate-specific membrane antigen expression. This combination enables the development of algorithms capable of distinguishing malignant uptake from physiologic activity across diverse anatomical regions and imaging conditions.

ToothFairy3 focuses on segmenting the Inferior Alveolar Canal (IAC) from Cone Beam Computed Tomography (CBCT) scans, a structure critical for safe surgical planning in dental and maxillofacial interventions [244]. The dataset comprises 3D CBCT volumes with voxel-level IAC annotations, facilitating the development of models that accurately localize and preserve this neurovascular bundle. The interactive challenge exclusively targets IAC segmentation to promote the development of interactive segmentation methods for detailed anatomical understanding and risk reduction during surgery.

TriALS targets automated liver lesion segmentation in portal-venous phase CT imaging, with the goal of enhancing lesion detection and delineation under clinically relevant conditions. The dataset contains multi-phase CT scans, with the portal-venous phase serving as the primary focus for interactive model development and evaluation. The challenge establishes a benchmark for robust and generalizable liver tumor segmentation, contributing to research toward clinically deployable AI-assisted liver imaging solutions.

MedSegFM introduces the first competition dedicated to interactive 3D biomedical image segmentation foundation models. Its goal is to develop and evaluate universal frameworks capable of handling diverse anatomical structures and imaging modalities while iteratively refining predictions through user interactions. The challenge provides a large-scale, well-curated dataset of over 200,000 3D image–mask pairs and employs comprehensive metrics to assess segmentation accuracy and interaction efficiency. Building on the CVPR 2024 "Segment Anything in Medical Images on Laptop" Challenge, the 2025 edition expands the focus to interactive foundation models. As part of this continuation, we co-organized the challenge and were responsible for designing the complete evaluation pipeline, including interaction simulation and evaluation metric implementation.



autoPET (PET/CT)

ToothFairy3 (CBCT)

TriALS (portal-venous CT)

MedSegFM (3D modalities)

Figure 46: Example of images from all co-organized challenges.

Examples for images from each challenge are given in Figure 46. This diversity poses a potential pitfall—if the interactive components are designed in a task-specific manner, the resulting insights may remain narrowly applicable to individual challenges. To mitigate the diversity of these challenges, we sought to unify all segmentation tasks under a common framework by applying general principles across: simulated interactions (Section 7.3.2), evaluation metrics (Section 7.3.3), and baseline models (Section 7.3.4). In the following, we describe how each of these three components was consolidated, ensuring that our findings extend beyond individual tasks. This cross-challenge perspective enables us to derive broader insights and best practices for the development of interactive segmentation models (Section 7.4.

7.3.2 *Interaction Simulation*

As discussed in Chapter 5, interaction simulations must be realistic to accurately reflect how a model would perform in clinical practice. The most reliable way to achieve this is by engaging directly with the intended end users by understanding who they are and how they will use the model. To this end, we conducted extensive discussions with each challenge organizer and designed a dedicated annotation protocol for each task.

Each annotation protocol consists of a fixed, well-defined sequence of steps describing how to annotate the challenge data using clicks. This ensures that the way we simulate data aligns closely with the annotation procedures that will be used later in practice when deploying the model. While each challenge has its own task-specific annotation protocol, which is necessary, given the diversity of the tasks, the overall framework for simulating and presenting these interactions to participants remains consistent across challenges. This unified simulation setup also enables a standardized post-challenge analysis of all tasks.

What is consistent across challenges? For all challenges, we implemented the *T2.1.2.2 Simulated Iterative Rule-based* taxonomy node introduced in Chapter 6. Iterative interactions capture how model performance evolves with an increasing number of user prompts, forming the core of human-in-the-loop evaluation. The rule-based component ensures that annotation protocols remain deterministic and reproducible in practice.

We also standardized the way interactions are provided to participants. As defined in Chapter 3, clicks are represented as discrete spatial coordinates. These can be stored as metadata associated with each image—for example, in a JavaScript Object Notation (*JSON*) file where the target serves as a key and the corresponding click coordinates are stored as an array of 3D points. For the *MedSegFM* challenge, we use a `.npz` dictionary format, which is methodologically equivalent to the *JSON* representation.

What differs between challenges? The annotation protocol itself. It is intentionally task-specific, as each challenge demands distinct interaction rules and annotation logic to best reflect the underlying clinical or technical objectives. Here, we briefly introduce the annotation protocol for each challenge and how we simulate it.

7.3.2.1 *autoPET Click Simulation*

Algorithm 1 illustrates the simulation of user clicks for lesion and background regions in PET images. For each image, 10 lesion components are randomly sampled, with initial clicks placed at their centers and additional boundary clicks added if fewer than 10 components exist. Random perturbations are applied to mimic realistic user variability. Background clicks are similarly sampled from high-uptake regions above the 99.75th percentile of PET intensities, providing a consistent and reproducible click distribution.

Algorithm 1 Simulation of Clicks for autoPET

```
Require: Ground-truth label $\mathcal{L}$, PET image $I$, perturbation noise $\sigma$
Ensure: Sets of simulated clicks $\mathcal{C}_{\text{lesion}}, \mathcal{C}_{\text{background}}$
function SIMULATE_AUTOPET_CLICKS($\mathcal{L}$)
    $\mathcal{C} \leftarrow \emptyset$; $\mathcal{L} \leftarrow$ select_random_10(components($\mathcal{L}$))      ▷ Select 10 random components
    for $l \in \mathcal{L}$ do
        $\mathcal{D} \leftarrow$ EDT($l$); $c \leftarrow$ argmax($\mathcal{D}$) $+ \mathcal{N}(0, \sigma)$      ▷ Center click with perturbation
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$
    end for
    while $|\mathcal{C}| < 10$ do      ▷ If $<$ 10 components, sample boundaries
        for $l \in \mathcal{L}$ do
            $\mathcal{D} \leftarrow$ EDT($l$); $c \leftarrow$ argmin($\mathcal{D}[\mathcal{L} > 0]$) $+ \mathcal{N}(0, \sigma)$      ▷ Boundary click with perturbation
            $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$; if $|\mathcal{C}| = 10$ then break
        end for
    end while
    return $\mathcal{C}$
end function
$\mathcal{C}_{\text{lesion}} \leftarrow$ SIMULATE_AUTOPET_CLICKS($\mathcal{L}$)      ▷ Clicks inside lesions
$\hat{\mathcal{L}} \leftarrow (I > P_{99.75}(I[\mathcal{L} = 0])) \odot (\mathcal{L} = 0)$      ▷ High-uptake non-lesion regions
$\mathcal{C}_{\text{background}} \leftarrow$ SIMULATE_AUTOPET_CLICKS($\hat{\mathcal{L}}$)      ▷ Background clicks
return $\mathcal{C}_{\text{lesion}}, \mathcal{C}_{\text{background}}$
```

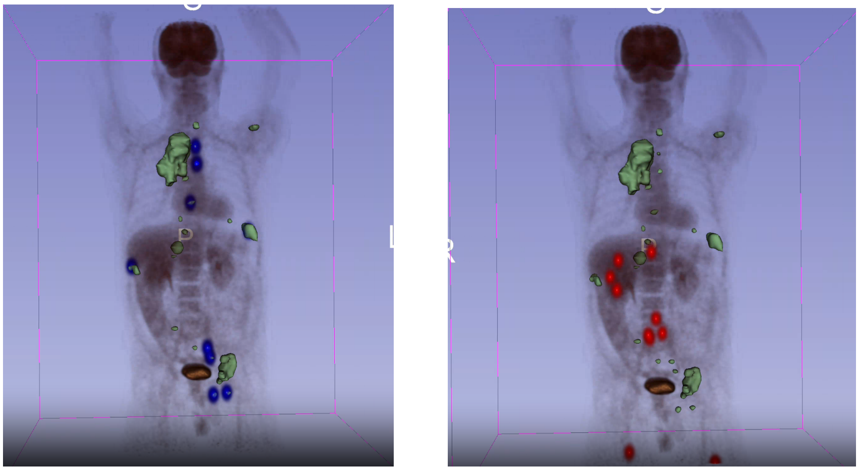


Figure 47: Lesion (left) and background (right) clicks simulated for the autoPET challenge. The green volumes are the ground-truth labels for the whole-body lesions.

7.3.2.2 *ToothFairy3 Click Simulation*

Algorithm 2 describes the simulation of clicks along the left and right Inferior Alveolar Canals (IAC) by sampling points in a slice-wise manner along the axial axis. For each selected axial slice, the geometric center of the mask is computed and then perturbed with small random offsets in the in-plane coordinates to emulate realistic variations in user annotation. This approach ensures that all simulated clicks remain within the target anatomical structure, while providing a spatially distributed set of points along the full extent of the canal. By repeating this procedure across multiple slices, the resulting click set captures both the longitudinal trajectory of the IAC and natural variability in user input.

Algorithm 2 Simulation of IAC Clicks for ToothFairy3

```
Require: Ground-truth label $\mathcal{L}$, number of clicks $N$, perturbation noise $\sigma$
Ensure: Sets of simulated clicks $\mathcal{C}_{\text{left}}, \mathcal{C}_{\text{right}}$
function SIMULATE_IAC_CLICKS($\mathcal{L}, N, \sigma$)
    $\mathcal{C} \leftarrow \emptyset$;
    axial_slices $\leftarrow$ sample_N_slices($\mathcal{L}$)        ▷ Sample N slices along axial axis
    for slice $\in$ axial_slices do
        $\mathcal{D} \leftarrow$ EDT(slice); $(x, y) \leftarrow$ argmax($\mathcal{D}$) $+ \mathcal{N}(0, \sigma)$        ▷ Center click with perturbation
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{(x, y, z)\}$
    end for
    return $\mathcal{C}$
end function
$\mathcal{C}_{\text{left}} \leftarrow$ SIMULATE_IAC_CLICKS($\mathcal{L}_{\text{Left\_IAC}}, N, \sigma$)        ▷ Left IAC clicks
$\mathcal{C}_{\text{right}} \leftarrow$ SIMULATE_IAC_CLICKS($\mathcal{L}_{\text{Right\_IAC}}, N, \sigma$)        ▷ Right IAC clicks
return $\mathcal{C}_{\text{left}}, \mathcal{C}_{\text{right}}$
```

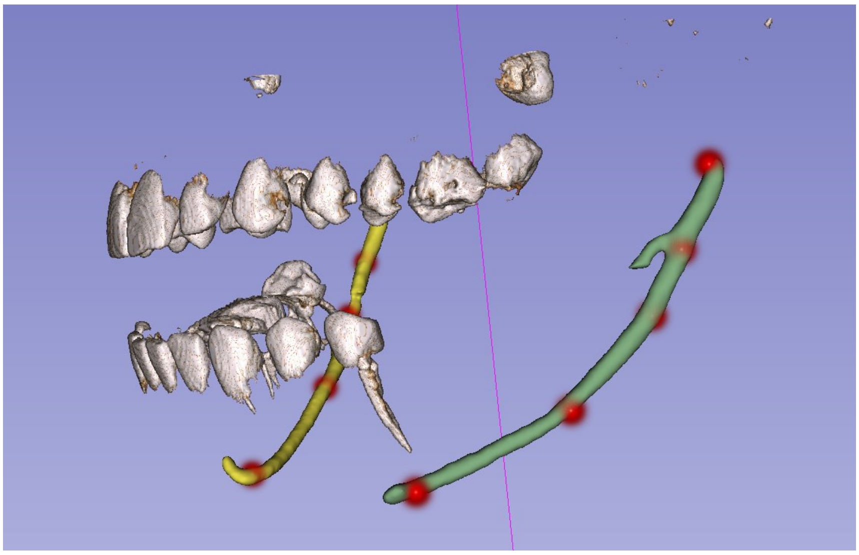


Figure 48: Left and right IAC simulated clicks (both in red) for the ToothFairy3 challenge.

7.3.2.3 *TriALS Click Simulation*

For the *TriALS* challenge, clicks are generated inside lesions and in the surrounding liver tissue. Lesion clicks are placed first at the center of each connected component in the label and supplemented with boundary clicks if fewer than 10 components are present. Background clicks are uniformly sampled within the liver mask outside the lesions, producing a set of simulated user interactions for both lesion and non-lesion regions. The liver mask is obtained with the *TotalSegmentator* model [291] and provided to the participants as metadata.

Algorithm 3 Simulation of Clicks for TriALS

```
Require: Ground-truth label $\mathcal{L}$, liver mask $\mathcal{M}$, perturbation noise $\sigma$
Ensure: Sets of simulated clicks $\mathcal{C}_{\text{lesion}}, \mathcal{C}_{\text{background}}$
function SIMULATE_LESION_CLICKS($\mathcal{L}$)
    $\mathcal{C} \leftarrow \emptyset$; $\mathcal{L} \leftarrow$ select_random_10(components($\mathcal{L}$))    ▷ Select 10 random lesion components
    for $l \in \mathcal{L}$ do
        $\mathcal{D} \leftarrow$ EDT($l$); $c \leftarrow$ argmax($\mathcal{D}$) $+ \mathcal{N}(0, \sigma)$    ▷ Center click with perturbation
        $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$
    end for
    while $|\mathcal{C}| < 10$ do    ▷ If $<$ 10 components, sample boundaries
        for $l \in \mathcal{L}$ do
            $c \leftarrow$ argmin(EDT($l$) $> 0$) $+ \mathcal{N}(0, \sigma)$    ▷ Boundary click with perturbation
            $\mathcal{C} \leftarrow \mathcal{C} \cup \{c\}$; if $|\mathcal{C}| = 10$ then break
        end for
    end while
    return $\mathcal{C}$
end function
function SIMULATE_BACKGROUND_CLICKS($\mathcal{M}, \mathcal{L}$)
    $\mathcal{C} \leftarrow$ uniform_sample($\mathcal{M} \cap (\mathcal{L} = 0), 10$)    ▷ Background clicks in liver
    return $\mathcal{C}$
end function
$\mathcal{C}_{\text{lesion}} \leftarrow$ SIMULATE_LESION_CLICKS($\mathcal{L}$)    ▷ Clicks inside lesions
$\mathcal{C}_{\text{background}} \leftarrow$ SIMULATE_BACKGROUND_CLICKS($\mathcal{M}, \mathcal{L}$)    ▷ Background clicks
return $\mathcal{C}_{\text{lesion}}, \mathcal{C}_{\text{background}}$
```

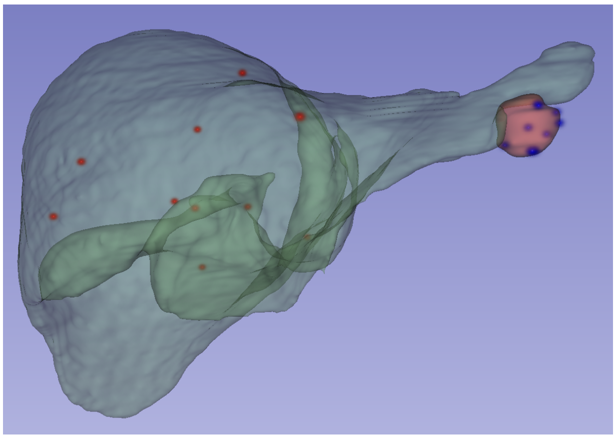


Figure 49: Lesion (blue) and background (red) clicks simulated for the TriALS challenge. The green volume is the liver pseudo-label and the red volume the ground-truth lesion mask.

7.3.2.4 *MedSegFM Click and Bounding Box Simulation*

The *MedSegFM* challenge also follows the *T2.1.2.2 Simulated Iterative Rule-based* taxonomy node, but we directly implement the robot user of Sakinis et al. [10], incorporating our efficiency optimizations from Section 4.2 to reduce simulation time by a factor of 20. The bounding boxes are also simulated by simply computing the maximum extent of the targets. We utilize a more generic robot user for this task as it covers 5 imaging modalities and over 100 segmentation targets, making it impossible to apply a task-informed interaction simulation method.

7.3.3 *Interactive Evaluation Metrics*

Evaluating interactive segmentation models requires metrics that explicitly capture how performance evolves as user interactions are added. As highlighted in Chapter 6, only a few prior works implement such interaction-aware metrics [209]. Some exceptions include user-centric measures like Dice@10 (Dice score after ten clicks) or full interaction curves showing performance as a function of the number of clicks. However, these are either too coarse (e.g., only the final Dice value) or too complex for consistent ranking across participants.

Our objective is to generalize existing challenge-specific metrics into interaction-aware versions that preserve their interpretability while enabling fair cross-method comparison. We achieve this by transforming any base metric $\texttt{metric}(\mathcal{L}, \mathcal{P})$, defined over a prediction $\mathcal{P}$ and ground truth $\mathcal{L}$, into two complementary interactive measures: (1) the **Area Under the Interaction Curve (AUC-metric)**, which summarizes performance across all interaction steps and reflects the overall efficiency of user guidance; and (2) the **Final metric value**, representing the segmentation quality after the full interaction process.

Together, these two metrics jointly capture both the trajectory and the endpoint of interaction-driven improvement, allowing consistent ranking of interactive methods across tasks. This evaluation framework is applied uniformly to **all four** challenges.

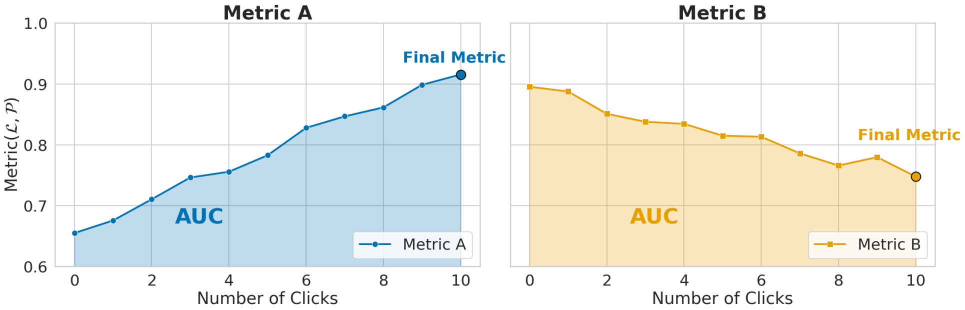


Figure 50: Examples of our interactive evaluation metrics. Our strategy can be applied to both positive metrics (A) and negative metrics (B) that need to be maximized or minimized respectively. The AUC- and Final metrics capture the nature of the interaction curve in two values.

7.3.4 *Interactive Baseline Models*

To provide participants with a common reference point for all interactive segmentation challenges, we employed the *SW-FastEdit* model introduced in Chapter 4 as a unified baseline. This model was selected for its efficient implementation of click-based refinement, making it a suitable foundation for diverse clinical tasks. For the *ToothFairy3* and *TriALS* challenges, we adapted the model by retraining it on each task's respective dataset while maintaining the same interactive inference pipeline as on the original *autoPET* data. This ensured that differences in baseline performance were attributable to the task characteristics rather than the model configuration itself. The resulting models serve both as reproducible reference points and as practical examples for participants to extend or adapt in their own submissions.

No additional baseline was trained for the *MedSegFM* challenge, as five external collaborators had already contributed comprehensive baseline implementations that fulfilled the same comparative purpose. Together, these baselines establish a starting reference point across all challenges and accelerate participant onboarding.

7.4 RESULTS FROM THE CHALLENGES: WHAT MAKES A GOOD INTERACTIVE MODEL?

Challenge	**Final Dice** **Best Interactive** **Model 2025**	**Final Dice** **Best Non-interactive** **Model 2022-2024**	**Improvement**
autoPET	71.4%	66.2%	5.2%
TriALS	75.8%	74.2%	1.6%
ToothFairy3	87.8%	79.5%	8.3%
MedSegFM	80.3%	–	–

Table 18: Final Dice scores of best models from the 2025 interactive challenges compared with the previous best results from their non-interactive counterparts (2023–2024).

Table 18 presents the final Dice scores achieved by the best models in our interactive challenges, alongside the previous best results obtained from their non-interactive counterparts. The interactive segmentation initiative has consistently advanced the performance boundaries established by non-interactive models across all evaluated challenges. Upon discovering this improvement, a natural question arises:

What makes a good interactive model?

Across all challenges, more than 200 submissions were evaluated during the development phase, leading to 30 successful final submissions in the test phase. Participating teams also provided technical reports and presented their approaches at our *CVPR 2025* and *MICCAI 2025* workshops. Through an in-depth analysis of these reports and presentations, we identified four key factors **(F1)–(F4)** that distinguish the winners, i.e., teams ranked in the top 3 positions for each challenge, from the rest. These factors are: **(F1)** iterative interactive training; **(F2)** small heatmaps $(\sigma < 5)$ or positional encoding (PE) as guidance signals; **(F3)** non-interactive pre-training; and **(F4)** local inference for corrective clicks. The difference between the percentage of winners and everyone else that has applied these strategies is seen in Figure 51.

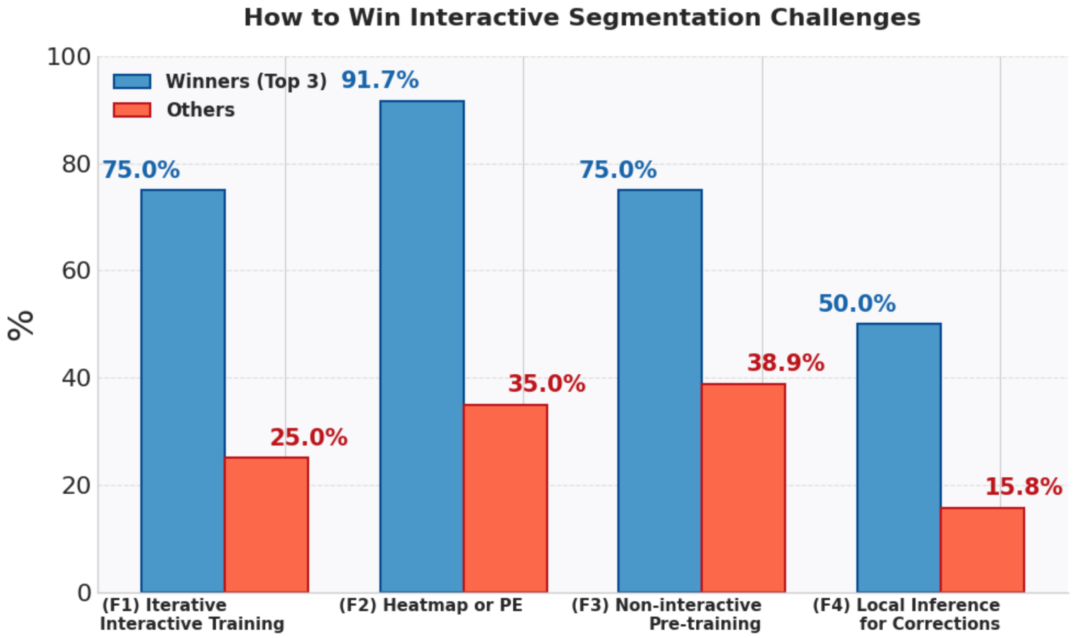


Figure 51: The four major contributing factors to win an interactive segmentation challenge.

(F1) Iterative Interactive Training. Three-quarters of the winning teams trained their models following the *T2.1.2. Simulated Iterative* taxonomy node, simulating multiple sequential interactions per image during training. In contrast, only 25% of non-winning teams adopted this strategy. Most non-winning methods merely combined the provided simulated interactions and treated clicks as just another imaging modality, without explicitly encouraging the model to improve as the number of interactions increased. As demonstrated in Chapter 4 through our ablation study with *SW-FastEdit*, this type of training regime can be detrimental to final model performance: it fails to effectively guide the model, ignores the sequential nature of interactions, and can even lead to overfitting due to the additional parameters introduced by extra channels. This likely explains why iterative interactive training is strongly associated with top-performing teams: it explicitly teaches the model to leverage sequential interactions, which aligns closely with the task requirements and improves overall robustness.

(F2) Small Heatmaps or Positional Encodings as Guidance Signals. Another factor distinguishing winning teams is the representation of clicks as guidance signals. Almost all winning solutions (91.7%) encode clicks as either small heatmaps ($\sigma < 5$) for U-Net backbones [127], or as positional encodings for transformer-based backbones like *VISTA 3D* [203]. This matches our findings in Chapter 3, where small-radius heatmaps and disks were shown to perform best for U-Net models [91]. The likely reason is that the local nature of small heatmaps ensures that sequential predictions remain stable: each new heatmap is self-contained and does not overwrite prior guidance. Small radii also provide precise overlap with ground-truth regions, which is particularly important for corrective clicks along edges, preventing large heatmaps from "overflowing" boundaries. This precision likely explains the prevalence of this strategy among top-performing teams.

(F3) Non-interactive Pre-training. Although the difference is smaller, most winning teams leveraged non-interactive pre-training to initialize their interactive models. This allows the model to start from a good local minimum, requiring only adaptation to the challenge dataset and learning to refine predictions through interactions. As confirmed in our ablation studies with *SW-FastEdit* (Chapter 4), non-interactive pre-training ensures that the initial predictions (with 0 clicks) are strong, which in turn improves AUC-based metrics that depend on the full prediction trajectory. This likely explains why pre-training provides an advantage: starting from a strong baseline allows interactive refinements to have a more consistent and measurable impact, improving final leaderboard performance.

(F4) Local Inference for Corrective Clicks. Although only 50% of winning teams applied this strategy, it was still notably more common than among non-winners (15.8%). Winning methods typically involve updating only a region around the new click or using zoom-in–zoom-out strategies to ensemble whole-image and cropped predictions, a technique with proven success in non-medical interactive tasks [216, 220]. While our experiments in Chapter 4 did not show a direct performance boost from local versus global inference, this approach reduces memory footprint and inference time. This efficiency advantage likely explains why top teams adopt it: it enables faster, more resource-conscious predictions without compromising accuracy, which can be crucial for practical deployment.

Challenger Winners. The *MedSegFM* winner, *iMedSTAM* [308], leverages simulated iterative user clicks during training to progressively refine segmentation predictions **(F1)**, encodes prompts as high-dimensional positional embeddings to guide the model's attention **(F2)**, and builds on *EfficientTAM*'s [309] pre-trained weights to provide robust and accurate initial segmentations **(F3)**. A distinctive feature of *iMedSTAM* is its bidirectional memory propagation across slices containing only the object of interest, which ensures spatial consistency and reduces error accumulation across the volume **(F4)**. The *autoPET* winner [310] similarly incorporates iterative click-based interactions **(F1)** and encodes user inputs as Gaussian heatmaps **(F2)**, while leveraging non-interactive pre-trained models to achieve strong initial predictions **(F3)**, enabling accurate segmentation with minimal user intervention. *ToothFairy3*'s winning model, *U-Mamba2* [311], integrates click positional embeddings through cross-attention into its mask decoder **(F2)** and benefits from pre-training on non-interactive dental CBCT datasets

(F3), facilitating efficient and reliable IAC segmentation. Finally, the *TriALS* CT liver lesion winner[9] encodes foreground and background clicks as Gaussian heatmaps **(F2)** within a residual U-Net [127, 136] enhanced with Response Fusion Attention [312], allowing the network to adaptively refine lesion delineation while maintaining standardized volumetric inference through the *nnU-Net* framework [136].

All of these factors **(F1)–(F4)** among the winning solutions are in agreement with our findings from our previous chapters, further confirming that our contributions to the foundations of interactive segmentation in Part ii have not only a theoretical implication but also emerge in practice when comparing the best approaches on various tasks.

7.5 CHAPTER CONCLUSION

Our interactive segmentation initiative has brought the discussion of research gaps in the field to the forefront of the community. By providing a unified framework for hosting interactive segmentation challenges and demonstrating its success across four independent events, we enable other competitions to adopt this structure, fostering exploration of new research directions and a broader range of segmentation tasks. The framework also allows for large-scale analysis of global trends in the field, helping identify the factors that define effective interactive segmentation models. Such insights are only possible when challenges are organized under consolidated principles, while still allowing task-specific flexibility such as dataset selection and annotation protocols.

We summarize the scientific impact of this chapter in three key contributions:

Contribution 1: A validated framework for transforming non-interactive challenges into interactive segmentation challenges by integrating simulated interactions, interactive metrics, and interactive baselines (*successfully applied in three MICCAI and one CVPR challenge*).

Contribution 2: Establishment of the first interactive segmentation community initiative, uniting challenge organizers and over 100 participants to collectively advance the field.

Contribution 3: Identification of four key factors that distinguish top-performing solutions and represent emerging best practices for interactive segmentation model development.

Our initiative represents the first coordinated community effort to address the research gaps in interactive segmentation. While the challenges have been highly successful, attracting strong engagement from the medical image analysis community and outperforming non-interactive approaches, the discussion remains largely centered around the challenge periods themselves.

[9] Slice & Dice: `https://github.com/Jess-Co-Del/TriALS2025`

To sustain progress, participants should be able to benchmark their methods not only during competitions but also on publicly available datasets with standardized evaluation protocols, complementing the challenge-based initiative. Given the growing number of open-access datasets, we ask:

How can we leverage public datasets to create a standardized interactive segmentation benchmark?

In the next chapter, we address this question by introducing *OmniMedSeg*—a large-scale interactive segmentation dataset and benchmark constructed from 166 open-licensed public datasets. It defines unified simulation and evaluation protocols inspired by the insights gained throughout this thesis.

8

STANDARDIZING MEDICAL IMAGE SEGMENTATION DATASETS

Medical image segmentation datasets are the foundation that drives the development of deep learning models across diverse segmentation tasks. As segmentation remains one of the central areas in the medical image analysis community, the availability of publicly annotated datasets has increased, enabling models to be trained on ever larger and more diverse data. However, due to the sensitivity of medical data, many datasets are restricted by licenses, limited to specific challenges, or require complex access procedures, such as signing data-use agreements or creating user accounts for tracking purposes. Moreover, medical imaging data is highly heterogeneous, existing in various formats that are often incompatible and difficult to use directly for model training or evaluation. In this chapter, we address these issues by unifying existing open-licensed datasets into *OmniMedSeg*—a large-scale, standardized benchmark for medical image segmentation. In Section 8.1, we discuss the motivation behind such a dataset and highlight the ongoing lack of *directly accessible* data resources for the community. Then, in Section 8.2, we introduce *OmniMedSeg*, and discuss how we collected and standardized public datasets, and extended them with interactive segmentation simulation and evaluation protocols to enable offline benchmarking of interactive segmentation models.

The contributions in this chapter and further analysis are *in submission* for peer review.

8.1 ON THE ACCESSIBILITY OF MEDICAL SEGMENTATION DATASETS

Public medical segmentation datasets have become increasingly common as the field continues to grow [92]. The rising popularity of segmentation challenges in recent years[1] has further contributed to the number of datasets available to the community [163]. However, when researchers begin developing their own methods for a specific task and explore existing data resources, they often discover that *public* does not necessarily mean *directly accessible* and *open*.

In many cases, medical datasets are only available during the active period of a challenge and become inaccessible once the challenge ends, effectively closing the data permanently [294]. Even when data remains technically public, access is often restricted: some datasets require users to sign user agreements and wait several days for approval [292, 293, 295]. Other platforms are less restrictive and only require account registration [296, 297], but this still poses

[1] https://miccai.org/index.php/special-interest-groups/challenges/miccai-registered-challenges/

limitations, as users must manually complete these steps in a browser, entering credentials and agreeing to terms, instead of accessing the data programmatically via an API or automated download link.

While such datasets can technically be classified as "public" since they can be obtained after completing certain procedures (e.g., emailing the authors, signing agreements, or creating accounts), they cannot be considered *directly accessible*. We consider directly accessible datasets as those hosted on platforms that provide a permanent and stable download link, allowing users to automatically retrieve the data through an API call or standard transfer protocols such as HTTP or FTP, without requiring additional authorization steps.

Moreover, even publicly available or directly accessible datasets often come with restrictive licenses that limit users' freedom to experiment or redistribute the data. For our purposes, we consider a dataset as *open-licensed* if its license permits redistribution, as this is essential for creating and sharing standardized versions of datasets. Conversely, datasets that prohibit redistribution, restrict use to specific purposes (e.g., challenge participation only), or impose similar constraints are not considered open-licensed.

For our dataset, *OmniMedSeg*, we include only datasets that are both *directly accessible* and *open-licensed*. This enables us to construct a framework capable of automatically generating standardized datasets using only the original download links, while also allowing us to redistribute these standardized resources freely within the research community.

8.2 OMNIMEDSEG: A LARGE-SCALE STANDARDIZED MEDICAL IMAGE SEGMENTATION DATASET

To establish the *OmniMedSeg* dataset, we follow a structured three-step pipeline: (1) comprehensive data collection; (2) conversion of all data into a standardized format; and (3) implementation of interactive segmentation simulation and evaluation protocols. The following sections describe each step individually — data collection in Section 8.2.1, data conversion in Section 8.2.2, and simulation and evaluation protocols in Section 8.2.3.

8.2.1 *Data Collection*

We adopted a systematic search strategy similar to that employed in our review in Chapter 6. However, instead of identifying studies on deep interactive segmentation, our focus here was on locating publicly available datasets for medical image segmentation.

We queried several prominent open-data repositories — including *Zenodo*, *Mendeley Data*, *Figshare*, and others (summarized in Table 19) — using targeted keyword combinations: [segmentation], [medical], and [image], in conjunction with modality-specific terms such as [fundus], [ultrasound], [MRI], [CT], [X-Ray], [dermoscopy], [endoscopy], [microscopy], and [OCT]. Our systematic search resulted in 314 segmentation datasets. Then, we manually reviewed the search results and included only datasets with: (1) directly accessible download links; (2) no prohibited redistribution; (3) no restricted use. This reduced the datasets to the final number of 166 datasets included in *OmniMedSeg*. A full list of the datasets is in the Appendix A.2.

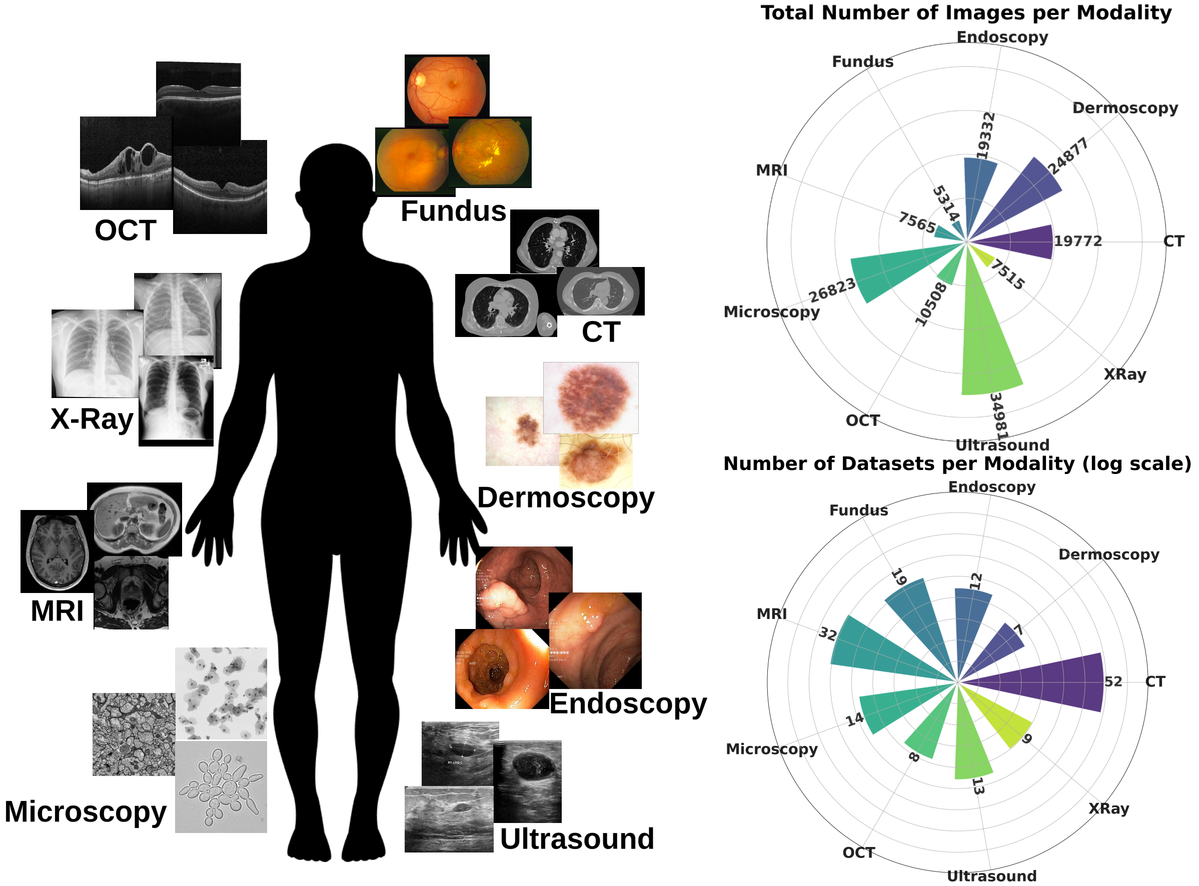


Figure 52: Overview of the *OmniMedSeg* dataset. We collected and standardized 166 open-licensed, directly accessible datasets from 9 imaging modalities.

Dataset Platform	Link
Figshare	`https://figshare.com/`
Mendeley Data	`https://data.mendeley.com/`
Zenodo	`https://zenodo.org/`
Grand Challenge	`https://grand-challenge.org/`
The Cancer Imaging Archive	`https://www.cancerimagingarchive.net/`
Kaggle	`https://www.kaggle.com/`

Table 19: List of all dataset platforms used in our systematic search for public datasets.

The distribution of the datasets among the nine imaging modalities can be seen in Figure 52. The CT and MRI domains dominate in terms of dataset availability, with 52 and 32 datasets, respectively, reflecting their widespread use and application. In contrast, modalities such as Dermoscopy and OCT remain underrepresented, each contributing fewer than ten datasets. When considering image volume, ultrasound and microscopy collectively account for the largest number of images (34,981 and 26,823). However, it is important to note that the CT and MRI number of images refers to volumes, each consisting of hundreds of slices.

8.2.2 *Conversion to a Standardized Format*

8.2.2.1 *On the Diversity of Medical Imaging Data*

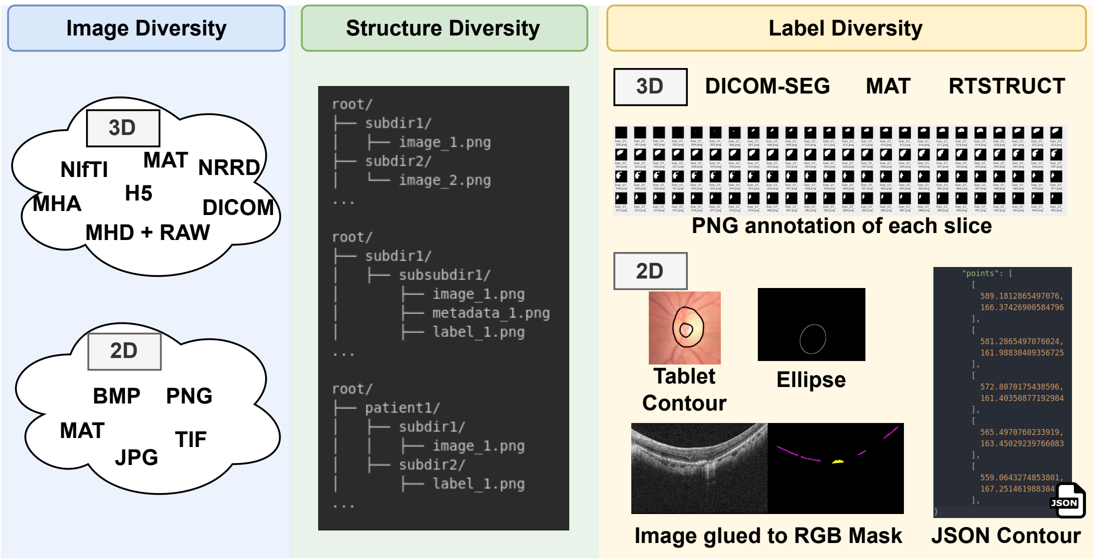
Medical imaging data exhibits substantial diversity across multiple dimensions, including image formats, directory structures, and label representations. This heterogeneity poses significant challenges for standardization and automation in data processing pipelines.

First, the variety of **image formats** is extensive. Common 2D formats include PNG and JPG, while legacy formats such as MAT also appear in certain datasets. For volumetric 3D data, formats such as NIfTI, DICOM, MHA, MHD+RAW, and H5 are prevalent, each carrying distinct metadata conventions and storage layouts. Second, **label formats** mirror this diversity. In 2D datasets, labels may be stored as binary masks or as text files describing contour coordinates. In 3D, labels are often encoded as binary NIfTI volumes, DICOM-SEG files, or RTSTRUCT objects. Third, the **dataset structure** itself can vary dramatically. Images may be stored in separate folders, grouped by patient, or combined with labels and metadata in a single directory. Such inconsistencies necessitate custom preprocessing for each dataset, often requiring significant engineering effort before training can begin. Figure 53 illustrates representative examples of this diversity in image formats, directory structures, and label representations.

During our exploration of publicly available datasets, we encountered several unconventional labeling practices. For example, some CT scan labels are provided as per-slice PNG masks. Other datasets use tablet annotations overlaid on RGB images, requiring machine learning–based extraction just to obtain usable binary masks. Additional cases include ellipses that must be filled to form binary labels, JSON contour files that require interpolation and rasterization, and RGB masks concatenated next to the image, where each color corresponds to a different class. Examples of each of these practices is given in Figure 53 (right).

To ensure consistent interaction simulation and evaluation across datasets, all data must first conform to a standardized format. This greatly improves usability, allowing datasets to be directly used for training and validation without requiring users to adapt to varying data structures or conventions. We eliminate the diversity of image and label formats and introduce a simple file structure to achieve this, as seen in Figure 54.

Previous efforts [201, 213, 262] standardized data through heavy normalization and filtering, simplifying usage but discarding valuable details such as metadata and labels with small structures. These approaches bias datasets toward large objects and break the dataset's integrity by



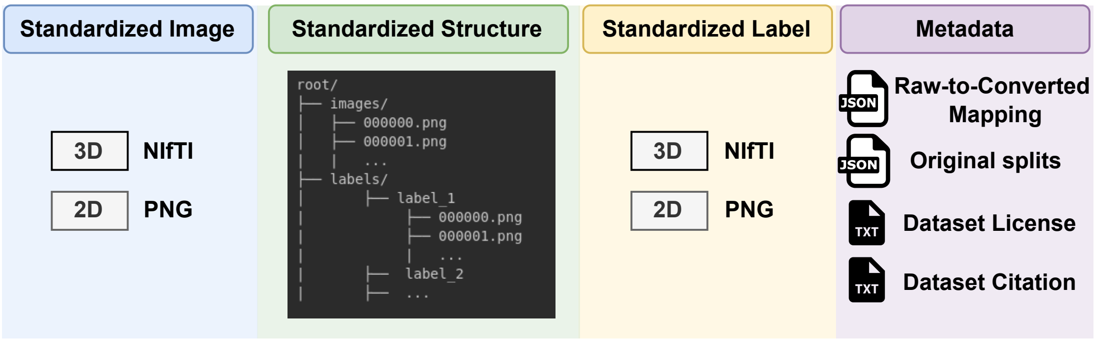
```
root/
├── subdir1/
│   ├── image_1.png
├── subdir2/
│   └── image_2.png
...

root/
├── subdir1/
│   ├── subsubdir1/
│       ├── image_1.png
│       ├── metadata_1.png
│       ├── label_1.png
...

root/
├── patient1/
│   ├── subdir1/
│   │   ├── image_1.png
│   ├── subdir2/
│       ├── label_1.png
...
```

```
    "points": [
      [
        589.1812865497076,
        166.37426900584796
      ],
      [
        581.2865497076024,
        161.98830409356725
      ],
      [
        572.8070175438596,
        161.40350877192984
      ],
      [
        565.4970760233919,
        163.45029239766083
      ],
      [
        559.0643274853801,
        167.251461988304
      ],
}
```

Figure 53: Examples illustrating the diversity of medical imaging datasets across image formats, directory structures, and label representations.

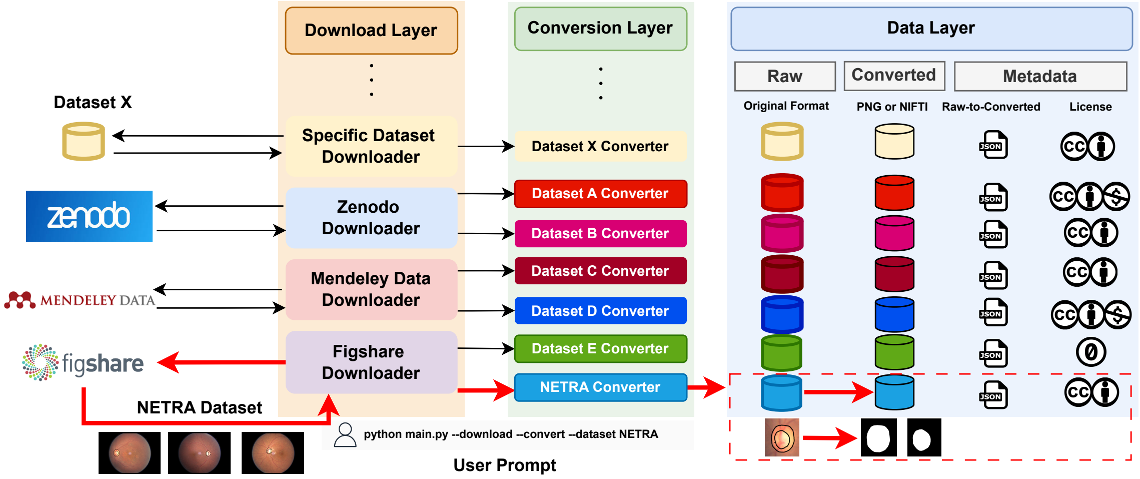


```
root/
├── images/
│   ├── 000000.png
│   ├── 000001.png
│   │   ...
├── labels/
│       ├── label_1
│           ├── 000000.png
│           ├── 000001.png
│           │   ...
│       ├──  label_2
│       ├──  ...
```

Figure 54: We use the NIfTI and PNG formats to store images and labels for all datasets in *OmniMedSeg*, and store a binary label for each class. We use a simple file structure, uniform for all datasets, and provide metadata: a JSON file that maps each converted image and label to the original file(s) used to create it; a JSON file indicating the original data splits if given; and the original data license and citation of the dataset.

discarding significant portions from it [120], e.g., all slices with small or no objects. In contrast, our work standardizes datasets while preserving the original data and labels, providing both a unified input format (as in Figure 54) and full access to the original files to ensure flexibility and access to all available clinical metadata that may be lost during the conversion process. We also provide a mapping of the converted images and labels to the original data to ensure that users can trace back the origin of each image and label in the converted datasets to the original files used to create them.

We designed *OmniMedSeg* with a modular architecture that allows the community to integrate additional datasets in the future without altering existing ones. The framework is organized into three core layers that ensure standardized and convenient access to all collected datasets: (1) the **Download Layer**, (2) the **Conversion Layer**, and (3) the **Data Layer**, illustrated in Figure 55.



```
python main.py --download --convert --dataset NETRA
```

Figure 55: Overview of the three layers behind our *OmniMedSeg* dataset. Here, the user requests the download and conversion of the NETRA dataset [307] (red arrows).

8.2.2.2 *Download Layer*

The download layer automates the retrieval of each of the 166 collected datasets through Application Programming Interface (API) calls or standard HTTP/FTP requests. We assign a dedicated *downloader* to each dataset platform (e.g., *Figshare* or *Zenodo*), which can be reused for all datasets hosted on that platform. For datasets stored on institutional or lab-specific servers, we implement custom downloaders. Each downloader operates independently, ensuring modularity and enabling seamless integration of new datasets without disrupting existing ones. This design allows users to access data effortlessly, without manually locating sources or performing multiple setup steps.

8.2.2.3 *Conversion Layer*

The conversion layer assigns a unique *converter* to each dataset to handle its specific format and structure. While converters are dataset-specific, common operations, such as DICOM-to-NIfTI conversion, are shared across implementations to avoid redundant code. Each converter transforms the dataset from its *raw format* into a *standardized format*: PNG for 2D images and NIfTI for 3D volumes. To extend the framework to new datasets, users simply need to implement one additional *converter* function, maintaining the consistency and modularity of *OmniMedSeg*.

8.2.2.4 *Data Layer*

The data layer manages both the converted and optional raw data. For each dataset, we provide a comprehensive JSON file mapping every converted image and label to its corresponding raw source files. For example, a single NIfTI file may be derived from multiple DICOM files, all of which are linked to the NIfTI filename in the JSON structure. This traceability allows users to reference original metadata, verify data integrity, or extract additional information if needed. Furthermore, we include the original license and citation information for every dataset, ensuring proper attribution and clarity on permissible use.

8.2.3 *Interactive Segmentation Evaluation and Simulation Protocols*

Given the standardized structure of *OmniMedSeg*, we can define interaction simulation and evaluation protocols in a dataset-agnostic manner. This abstraction allows the same protocol to be applied consistently across all 166 included datasets, as well as to future datasets added to *OmniMedSeg*. The central challenge when designing such protocols is to ensure they generalize effectively to datasets with diverse characteristics while remaining dataset-agnostic.

8.2.3.1 *Standardized Evaluation Protocols*

For the **interactive evaluation**, our prior experience co-organizing multiple public challenges in Chapter 7 provides a practical solution. Any non-interactive metric can be transformed into an interactive one by analyzing its evolution across interaction steps, specifically by computing the *Area Under the Curve (AUC)* and the *Final* metric value, which together characterize the interaction curve. This approach has been validated in four independent challenges to rank participants' approaches, suggesting that it would generalize well to other tasks.

In *OmniMedSeg*, we formalize this approach by implementing the *AUC* and *Final* metrics as lightweight wrappers around an abstract `Metric` class. Users can simply insert their metric implementation within this abstract class. Given a sequence of metric values collected over successive interaction steps, the AUC is computed via trapezoidal numerical integration, while the Final metric simply corresponds to the last recorded value. This modular design allows users to transform any existing metric (e.g., Dice score) into an interactive version with minimal effort by using the provided wrapper (as in Algorithm 4). Furthermore, *OmniMedSeg* includes reference implementations for the most common metrics identified in our systematic review (see Chapter 6), including Dice, IoU, Hausdorff Distance, among others. The AUC and Final metrics are metric-agnostic and can be applied to any user-defined implementation seamlessly.

Algorithm 4 Interactive Metric Computation in *OmniMedSeg*

```
abstract class Metric():
    def compute(preds, gts):
        # User-defined metric implementation here
        return metric_values
class InteractiveMetric():
    def compute_interactive(metric_values):
        AUC ← trapezoidal_integration(metric_values)
        Final ← metric_values[-1]
        return {AUC, Final}
Usage Example:
user_metric = Metric()
metric_values = user_metric.compute(predictions, ground_truths)
results = InteractiveMetric().compute_interactive(metric_values)
print(results)
```

8.2.3.2 *Standardized Simulation Protocols*

Defining a standardized **interaction simulation protocol**, i.e., standardizing the behavior of a robot user, is inherently challenging, as interaction dynamics often depend on dataset- and task-specific factors. Nevertheless, we observed that most task-specific robot users can be expressed as variations of a more general abstraction. For instance, in both the *autoPET* and *TriALS* challenges, interactions primarily consist of center and boundary clicks for lesions, differing only in how background points are sampled: from liver tissue in *TriALS* or from non-lesion, high-uptake regions in *autoPET*.

At a fundamental level, a robot user can be understood as a function that selects interaction points based on an **eligibility mask** $\mathcal{M}$—a binary mask indicating valid sampling regions. The eligibility mask can represent various things, such as ground-truth labels $\mathcal{L}$, model errors $\mathcal{L} \neq Y$ (where Y is the model prediction), or other logical combinations derived from the image I, label $\mathcal{L}$, and prediction Y. Formally, $\mathcal{M}$ can be expressed as the output of a function $f(I, \mathcal{L}, Y)$, which encodes task-specific eligibility criteria. This function f can even define the eligibility mask $\mathcal{M}$ to contain regions *outside* the ground-truth mask $\mathcal{L}$ to emulate a degree of inter-annotator disagreement, as we did in Chapter 5.

For example, in *autoPET*, background clicks are sampled from an eligibility mask $\mathcal{M}$ containing only high-uptake, non-lesion voxels. In *TriALS*, $\mathcal{M}$ corresponds to the liver region predicted by TotalSegmentator [291]. The key insight is that robot user design only needs to operate on this abstract concept of an eligibility mask, and the transformation function f adapts the simulation to the specific dataset or task. Thus, while implementations of f may vary, the underlying framework remains task-agnostic. Within this abstraction, we define standardized protocols for four interaction types: clicks, scribbles, bounding boxes, and polygon vertices, each with multiple variations inspired by the robot users we found during our review in Chapter 6. Users can select from these protocols depending on their use case or employ identical settings with prior work to ensure comparability.

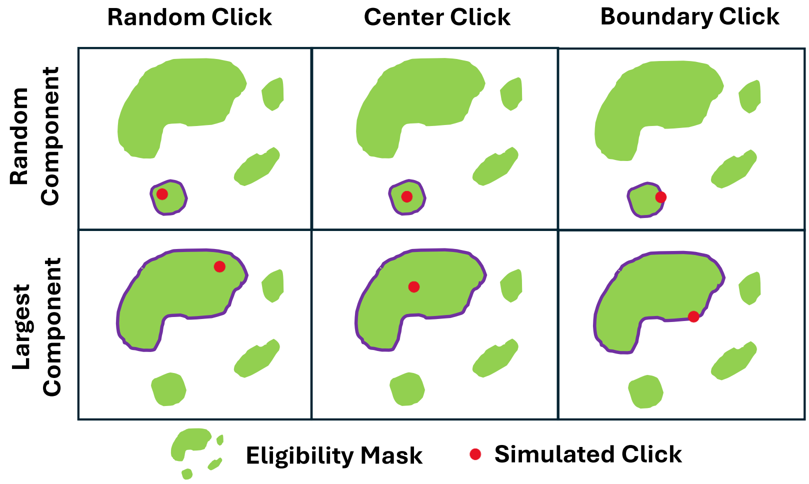


Figure 56: Standardized click-based robot users in *OmniMedSeg*.

Clicks. Given an eligibility mask $\mathcal{M}$, we simulate clicks following six annotation styles. The simulation proceeds in two stages: (1) selecting either the largest or a random connected component in $\mathcal{M}$, and (2) placing the click either at the center, boundary, or a random location within the component. These two stages yield six robot users with distinct annotation behaviors. Center and boundary clicks are generated by computing the Euclidean Distance Transform (EDT) of $\mathcal{M}$ and selecting from either its maximum or minimum points, i.e., $\text{argmax}(\mathrm{EDT}(\mathcal{M}))$ or $\text{argmin}(\mathrm{EDT}(\mathcal{M}) \odot \mathcal{M})$. Representative examples are shown in Figure 56.

Scribbles. We implement six types of analogous robot users for scribble-based interactions, following the established ScribblePrompt methodology [202]. For centerline scribbles, we extract the skeleton of $\mathcal{M}$, apply a random mask to fragment the structure, and introduce a deformation field to simulate user variability in stroke thickness and continuity. The resulting scribble is then masked with $\mathcal{M}$ to ensure it remains confined to valid regions. Boundary scribbles are generated by Gaussian-smoothing $\mathcal{M}$, thresholding to isolate edge regions, and subsequently applying the same random masking and deformation process. Random scribbles, consistent with *ScribblePrompt* [202], are produced by sampling random points within $\mathcal{M}$, connecting them with lines, and applying the same fragmentation and deformation steps as above. Illustrative examples are shown in Figure 57.

Bounding Boxes. We provide three robot user variants for bounding boxes. The *perfect user* always draws an exact ground-truth bounding box. The *careful user* expands each dimension uniformly by 5–10%, simulating conservative over-annotation. Finally, the *sloppy user*

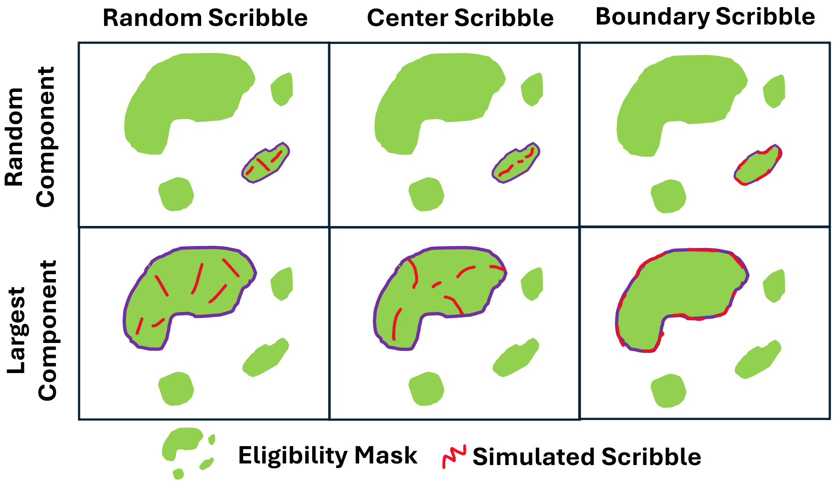


Figure 57: Standardized scribble-based robot users in *OmniMedSeg*.

introduces asymmetric perturbations of 5–10% in random directions along each dimension, mimicking inconsistent or imprecise annotation behavior.

Polygon Vertices. Polygon-based interactions are implemented only for 2D data. Given a user-specified sampling budget N, N vertices are sampled uniformly over the boundary of the eligibility mask $\mathcal{M}$, forming a polygonal contour representation of the target region.

All of these robot users are integrated within the *OmniMedSeg* benchmark and can operate in both iterative and non-iterative settings, covering all areas in the interactive segmentation taxonomy [209]. We provide reference implementations and example scripts with a dummy model to demonstrate evaluation using any robot user, allowing users to simply insert their models by replacing the dummy model in the provided framework. Our standardized evaluation and simulation protocols unify the diverse datasets and interaction types. By abstracting user behavior through the concept of an eligibility mask $\mathcal{M}$ and formalizing interaction primitives for clicks, scribbles, bounding boxes, and polygons, *OmniMedSeg* bridges the gap between dataset-specific implementations and generalizable evaluation pipelines, laying the foundation for fair, consistent, and extensible research in interactive medical image segmentation.

8.3 CHAPTER CONCLUSION

We present *OmniMedSeg* — a large-scale, multimodal dataset that unifies 166 openly licensed datasets spanning nine imaging modalities. *OmniMedSeg* establishes a standardized foundation for training and evaluating both non-interactive and interactive segmentation methods, promoting accessibility, reproducibility, and fair comparison within the medical image segmentation community. The framework is designed to be transparent and extensible, built upon generic, dataset-agnostic concepts that can be uniformly applied across all included datasets and easily expanded with new contributions. Its open-source nature encourages community-driven development, enabling researchers to extend, refine, and maintain the resource collaboratively. Beyond data aggregation, *OmniMedSeg* provides standardized simulation and evaluation protocols for interactive segmentation, ensuring consistent benchmarking conditions. These components are designed to evolve in tandem with the community, supporting future extensions and methodological advances in interactive medical image segmentation.

We summarize the scientific impact of this chapter in three key contributions:

Contribution 1: *OmniMedSeg* - a comprehensive multi-modal dataset. It introduces a standardized, large-scale resource that unifies 166 publicly available datasets across nine imaging modalities, enabling consistent data access and integration across the medical imaging domain.

Contribution 2: A standardized interactive evaluation and simulation framework built on top of *OmniMedSeg*, ensuring fair, reproducible, and comparable benchmarking conditions for both existing and future interactive segmentation methods.

Contribution 3: An extensible and community-driven design. *OmniMedSeg* is designed as a living framework rather than a fixed solution. Its modular and extensible architecture supports continuous integration of new datasets, modalities, and protocols, allowing the community to expand and refine the resource collaboratively over time.

OmniMedSeg represents the first unified, large-scale effort to standardize medical image segmentation data, simulation, and evaluation under a single open framework. By standardizing heterogeneous datasets and formalizing reproducible interactive evaluation and simulation protocols, it provides a robust foundation for developing and benchmarking both interactive and non-interactive segmentation methods. More importantly, its open and modular design ensures that *OmniMedSeg* will continue to evolve alongside the field, fostering collaboration, transparency, and sustained progress in medical image analysis research.

Part IV

INSIGHTS

9 IMPACT ON THE FIELD

This thesis has contributed to advancing the field of medical interactive segmentation across multiple dimensions. In Part ii, we examine three fundamental components of interactive segmentation: *representation* (Chapter 3), *efficiency* (Chapter 4), and *simulation* (Chapter 5); and propose novel methods that address key challenges within each. In Part iii, we establish a comprehensive taxonomy of the field (Chapter 6), foster community engagement through the co-organization of public segmentation challenges (Chapter 7), and introduce a large-scale standardized dataset (Chapter 8). This chapter revisits these core contributions, emphasizing their broader impact from four perspectives: new research directions, new models, new datasets, and new communities.

9.1 NEW RESEARCH DIRECTIONS

INTERACTION REPRESENTATION. Our comparative analysis of interaction representations in Section 3.2 demonstrates that the choice of a guidance signal substantially influences model performance in terms of segmentation quality, efficiency, and responsiveness to user input. We also found the choice of the guidance signal to be crucial to win interactive segmentation challenges in Section 7.4. These findings highlight the interaction representation as a fundamental design axis in interactive segmentation, opening a new, unexplored research direction to investigate and formalize the effects of different interaction types, such as scribbles, bounding boxes, and polygon vertices, on model performance, generalization, and task adaptability.

MODEL-AGNOSTIC EFFICIENT INTERACTIVE SEGMENTATION. We propose two model-agnostic strategies that improve the computational efficiency of *any* interactive segmentation model: localized inference via sliding-window processing (Section 4.1) and accelerated distance transform computation (Section 4.2). Rather than optimizing a specific architecture, our methods target the broader paradigm of interactive segmentation itself by introducing efficiency improvements that can be integrated into diverse interactive models. These contributions emphasize efficiency as a fundamental dimension of interactivity and open a new direction for developing *model-agnostic* efficiency methods that enable scalable, responsive, and real-time performance, in contrast to prior work limited to specific models and tasks.

REALISTIC INTERACTION SIMULATION. In Section 5.2, we reveal how existing robot user protocols often produce overly optimistic performance estimates when evaluating interactive models. This observation identifies realistic interaction simulation as an important and novel research direction, as previous approaches in the field still adopt overly simplistic simulation approaches. We found that factors such as inter-annotator disagreement and human variability capture important aspects of realism. However, they represent only a part of a broader landscape of human interactions. Our work opens a new research direction focused on formalizing and systematizing *realistic* human–model interaction simulation, enabling the community to explore how factors such as cognitive biases, annotation strategies, and contextual uncertainty influence the realism of interaction simulations.

COMMUNITY-DRIVEN INTERACTIVE SEGMENTATION. During our review in Section 3.2, we found that the interactive segmentation field remains fragmented and lacks the community infrastructure needed for sustained, coordinated development. To mitigate this issue, in Chapters 7 and 8, we demonstrate that public challenges and standardized datasets can drive community-driven progress by fostering transparency, reproducibility, and shared evaluation practices. Our analysis of the challenges we co-organized reveals common characteristics among top-performing teams (Section 7.4), outlining emerging best practices for interactive models. Moreover, the *OmniMedSeg* dataset we introduce in Section 8.2.1 offers a scalable and modular infrastructure designed to evolve alongside the field and its community. Together, these contributions open a new research direction toward *community-driven* interactive segmentation, where progress is shaped through open collaboration and collective insight.

9.2 NEW METHODS AND FRAMEWORKS

We introduce novel practical methods that advance three key dimensions of interactive segmentation: 1) *representation*, 2) *efficiency*, and 3) *realistic simulation*:

1. In Section 3.2.4, we propose our adaptive Gaussian heatmaps as a novel guidance signal that dynamically adjusts its radius based on underlying image features. Our approach unifies geodesic distances with iterative interactions for the first time and demonstrates superior performance across two segmentation tasks.

2. In Section 4.1, we present *SW-FastEdit*, a model that employs localized inference to maintain constant inference speed and memory usage while ensuring rapid responsiveness to user input. This enables efficient application to volumes of arbitrary size, overcoming the scalability and memory limitations that have constrained previous interactive approaches.

3. In Section 5.2, we introduce a novel robot user that integrates inter-annotator disagreement into the simulation process, yielding significantly more realistic and representative evaluations. We validate our robot user through two independent user studies with medical annotators.

Beyond practical approaches, in Section 6.1, we also establish a *theoretical* framework for interactive segmentation that brings conceptual clarity and structural coherence to the field. Our framework can categorize any interactive model within a formal taxonomy by answering a systematic set of questions, provides rigorous definitions for core concepts such as the *guidance signal* and the *robot user*, and identifies fundamental research gaps in the field by analyzing trends among existing methods within a coherent and extensible structure.

9.3 NEW COMMUNITY

In Chapter 7, we introduce our *Interactive Segmentation Initiative*, a community-driven effort to standardize the field of medical interactive segmentation by co-organizing challenges in collaboration with research groups worldwide. Our initiative unites diverse groups of people: from challenge organizers with domain expertise to participants developing competing methods. Together, they form a collaborative ecosystem that collectively identifies best practices, exposes methodological strengths and weaknesses, and drives progress through competitive yet cooperative experimentation. Initially hosted at *MICCAI 2025* and *CVPR 2025*, the initiative is designed for long-term growth, with additional challenges expected to join in future iterations and existing collaborations expanding their scope and datasets. By fostering global coordination and continuous dialogue, this community elevates research questions to a broader scale and advances the pursuit of clinically meaningful, reproducible, and scalable solutions in medical interactive segmentation.

9.4 NEW DATASET

In Chapter 8, we consolidate the first large-scale, open-licensed, and directly accessible collection of medical image segmentation datasets that unifies data accessibility with standardization - *OmniMedSeg*. Our dataset preserves both the original imaging data and its associated meta- and clinical information, while also providing a standardized version of 166 public datasets to enable researchers to develop and evaluate their methods under identical conditions. The collection spans nine imaging modalities and offers a transparent, modular framework for downloading, converting, and managing datasets that facilitates long-term extensibility and enables the community to add additional datasets in a consistent format. Because all datasets share the same structure, we are able to apply uniform interaction simulation and evaluation protocols across the entire collection. These protocols ensure that interactive methods can be compared using the exact same data and experimental conditions, leading to more reproducible, interpretable, and fair benchmarking in interactive medical image segmentation.

10 OPEN RESEARCH QUESTIONS

This thesis has introduced practical (Chapters 3, 4, 5), theoretical (Chapter 6), and community-based (Chapters 7, 8) contributions to the medical interactive segmentation paradigm. The overarching aim of these efforts is to *standardize* the field and mitigate its fragmented nature. By establishing this foundation, our work enables the exploration of new research questions that were previously inaccessible due to the lack of common frameworks and benchmarks. In the following, we outline the most relevant open research questions that can now be addressed based on the foundations laid in this thesis.

10.1 FOUNDATION OR SPECIALIST INTERACTIVE SEGMENTATION MODELS?

Foundation models have achieved remarkable performance across general computer vision tasks in recent years [137, 298, 299]. This trend is rapidly extending to medical image analysis, where many approaches have adapted the Segment Anything Model (SAM) [137], released in 2023, for medical applications [93, 118–120]. At the time of writing, there are over 345 medical adaptations of SAM[1] , highlighting its wide adoption. Thanks to its generalization and zero-shot capabilities, SAM-based methods often evaluate across a large number of tasks, with some approaches tested on more than 30 public medical datasets [119, 120].

However, as discussed in our review in Chapter 6, most SAM-based methods are compared only against SAM itself rather than other interactive models [209]. This raises the question:

Are foundation models truly better than specialist models?

Answering this question is challenging because foundation models exist at different granularities [300]: some are fully general, while others are modality-specific (e.g., an Ultrasound foundation model). Specialist models, by contrast, are tailored to particular tasks or targets (e.g., a lung cancer segmentation model). Evaluating whether general foundation models offer tangible advantages over specialist models requires direct, systematic comparison.

Our large-scale *OmniMedSeg* dataset provides an ideal platform for such analyses. By unifying diverse datasets into a consistent format, it enables direct comparisons between models, simplifies experiment design, and ensures reproducibility, making it possible to rigorously assess the trade-offs between foundation and specialist approaches.

[1] https://github.com/YichiZhang98/SAM4MIS

10.2 TEXT-BASED INTERACTIONS

With the rise of large language models (LLMs) [301–303] and their growing adoption in medical applications [304–306], integrating the language domain into interactive segmentation has become increasingly relevant. At the time of our systematic review in Chapter 6, text-based medical interactive segmentation was still an emerging area, with no studies included in the review. Nevertheless, text-based interactions represent a promising direction for the field, enabling users to specify segmentation targets via natural language queries.

Recent methods leverage text in two main ways. Some approaches use text *directly* as an additional input to the interactive model, typically employing image-text pairs for contrastive pre-training to align text and image embeddings corresponding to the same target [183–185]. Other approaches use text *indirectly* via visual grounding: a text query is processed by a detection or grounding model to generate a bounding box, which is then used as an interaction cue for the segmentation model [181, 182, 233].

Text-based interactions are expanding the scope of interactive segmentation beyond traditional image-based inputs, opening avenues for incorporating speech and other modalities. Given that language-based cues differ fundamentally from vision-based interactions, an important open question remains:

How can we effectively leverage text to guide interactive segmentation models?

10.3 EXPLORATION OF INTERACTIVE SEGMENTATION EVALUATION METRICS

In this thesis, we introduce several novel metrics for evaluating interactive segmentation models, ranging from metrics for guidance signal effectiveness (Chapter 3) and user shift between real and simulated robot users (Chapter 5), to interactive extensions of standard metrics used in challenge rankings (Chapter 7) and incorporated in *OmniMedSeg* (Chapter 8).

A central insight in designing these metrics was the importance of expert feedback. All metrics were iteratively refined with input from medical annotators and challenge co-organizers, ensuring evaluations align with the needs of end-users.

Nonetheless, it remains an open question how to broaden this discussion to the wider community. One possibility is to emulate the *Metrics Reloaded* initiative [165], where a consortium of experts define guidelines for metric selection. A similar effort in the interactive domain could be even more impactful, as human interaction is integral to these models. Our *Interactive Segmentation Initiative* (Chapter 7) provides a starting point to assemble the expertise needed for such a consortium to answer the following research question:

How can the community define evaluation metrics for interactive segmentation models?

Part V

APPENDIX

A

APPENDIX

A.1 ADDITIONAL EXPLICIT GUIDANCE SIGNALS

Here, we define the rest of the explicit guidance signals we have found in literature in our *TPAMI 2024* [209] publication.

Location Priors, as proposed by Sun et al. [3], incorporate both the Manhattan distance and the information about crossed edges detected by a Canny edge detector [123]. The location prior assigns an initial intensity value of 255 to the click location, denoted as $c = [c_x, c_y]$, and decreases this value by 1 for each vertical or horizontal step taken for the other pixels $v = [v_x, v_y]$ in the signal. Furthermore, when a step crosses a detected edge in $\mathrm{Canny}(I)(i,j)$ at a pixel $[i,j]$, the intensity value decreases by an additional 10. This guidance signal combines the notion of distance with the presence of edges to provide a comprehensive measure for location estimation as defined in Equation 29. It is also originally defined only for 2D images.

$$\mathrm{LP}(v) = \max\left(0, 255 - \sum_{i=\min(c_x,v_x)}^{\max(c_x,v_x)} \sum_{j=\min(c_y,v_y)}^{\max(c_y,v_y)} \begin{cases} 10 & \text{if } \mathrm{Canny}(I)(i,j) == 1 \\ 1 & \text{otherwise} \end{cases}\right) \tag{29}$$

Attraction Field Weight Maps (AFWM), as introduced in [14], draw inspiration from the attraction field generated by punctual electric charges of opposite values. It utilizes unitary gradient fields, denoted as $\nabla S_i(v)$, over voxels/pixels $v = [v_x, v_y]$, which are centered around two clicks, namely $c_1 = [c_{1x}, c_{1y}]$ and $c_2 = [c_{2x}, c_{2y}]$. These gradient fields exhibit higher values between the clicks, indicating their significance for the segmentation process. The hyperparameter $p \in \mathbb{R}$ controls the decay of the vectors' magnitude, and the signal is defined in Equations 30 and 31:

$$\mathrm{AFWM}(v, c_1, c_2) = \frac{\nabla S_1(v)}{|\nabla S_1(v)|^p} - \frac{\nabla S_2(v)}{|\nabla S_2(v)|^p} \tag{30}$$

$$\nabla S_i(v) = \frac{2(v_x - c_{ix}) + 2(v_y - c_{iy}) + (v_z - c_{iz})}{2||v - c_i||} \tag{31}$$

Positional Encodings (PE) are encoded as point-based positional embeddings, and are a fundamental block of vision transformer (ViT) models, such as the Segment Anything Model (SAM) [137], that operates on 2D images. Each click $c = [c_x, c_y]$ is encoded as:

$$E(c) = \begin{bmatrix} \sin(\omega_1 c_x), \cos(\omega_1 c_x), \ldots, \sin(\omega_K c_x), \cos(\omega_K c_x), \\ \sin(\omega_1 c_y), \cos(\omega_1 c_y), \ldots, \sin(\omega_K c_y), \cos(\omega_K c_y) \end{bmatrix} \in \mathbb{R}^{4K}, \tag{32}$$

where $\omega_k = 1/10000^{2k/d}$ are fixed frequency terms, with the hyperparameters $K \in \mathbb{N}$ defining the number of frequency bands, and $d \in \mathbb{N}$ is the dimensionality of the embedding. The complete embedding is then obtained by adding this positional encoding to a learned prompt embedding $\mathbf{e} \in \mathbb{R}^{4K}$:

$$\mathrm{PE}(c) = E(c) + \mathbf{e}, \tag{33}$$

with the overall click guidance signal is aggregated as:

$$\mathrm{PE}(\mathcal{C}) = \sum_{c \in \mathcal{C}} \mathrm{PE}(c) \tag{34}$$

Scribble-based Explicit Guidance Signals have a significant overlap with click-based signals, as a set of clicks $\mathcal{C}$ is, according to Equation 1 a subtype of scribbles. Hence, common explicit signals for scribbles, such as Gaussian heatmaps, Euclidean maps, and Geodesic maps, can be formally defined with the click-based Equations 4, 7, and 8, respectively. However, when intuitively thinking about scribbles, we tend to think about *connected* lines or curves, rather than isolated points in the image. Hence, there are some scribble-specific guidance signals seen in prior work, such as B-splines and scribble expansion methods [41].

B-Splines [272] transform a scribble $\mathcal{S} = \{p_1, \ldots, p_N\}$ into a continuous curve $\mathbf{B}(t)$ by treating the scribble points (or a subset) as control points and blending them smoothly via the B-spline basis functions $N_{i,j}(t)$, which determine the influence of each point on the curve. This approach allows the creation of smooth, visually coherent guidance curves from a sparse set of user-defined points, capturing both the global shape and local variations of the scribble. The explicit guidance signal is then computed as the binary curve $\mathcal{G}(v)$, as defined in Equation 35.

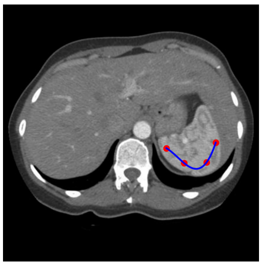


Figure 58: A B-spline scribble.

$$\begin{aligned}
\mathbf{B}(t) &= \sum_{i=1}^{N} N_{i,j}(t)\, p_i, \quad t \in [0,1], \\
N_{i,0}(t) &= \begin{cases} 1, & t_i \leqslant t < t_{i+1}, \\ 0, & \text{otherwise}, \end{cases} \\
N_{i,j}(t) &= \frac{t - t_i}{t_{i+j} - t_j} N_{i,j-1}(t) + \frac{t_{i+j+1} - t}{t_{i+j+1} - t_{i+1}} N_{i+1,j-1}(t), \\
\mathcal{G}(v) &= \begin{cases} 1, & v \in \{\mathbf{B}(t) \mid t \in [0,1]\}, \\ 0, & \text{otherwise}. \end{cases}
\end{aligned} \tag{35}$$

Scribble Expansion (SE) guidance signals [41] aim to increase the context of the scribble points $\mathcal{S} = \{p_1, \ldots, p_N\}$ by expanding each scribble using region growing [273]. For each voxel/pixel v, the expansion proceeds from the scribble points and includes neighboring voxels/pixels if the local intensity difference is below a threshold $T \in \mathbb{N}$. Formally, the guidance signal is defined as:

$$\mathrm{SE}(v, T) = \begin{cases} 1, & \exists\, p_i \in \mathcal{S} \text{ such that } v \text{ is reachable from } p_i \text{ with } |I(v) - I(p_i)| < T, \\ 0, & \text{otherwise,} \end{cases} \tag{36}$$

where $I(v)$ denotes the intensity at pixel v and the "reachable" condition encodes the standard region growing criterion: the expansion stops whenever the intensity difference exceeds T.

Identity Encoding (id) encodes the identity function, i.e., simply encoding the scribbles directly in a binary mask as defined in Equation 37. As this is the simplest way to implement scribbles, it is adopted by 61% of the scribble-based explicit guidance signals [209].

$$\mathrm{id}(v, \mathcal{S}) = \begin{cases} 1, & v \in \mathcal{S}, \\ 0, & \text{otherwise,} \end{cases} \tag{37}$$

Bounding Box-based Explicit Guidance Signals are not as common as click- or scribble-based signals, as most prior work uses the boxes to crop the image as an implicit guidance signal [209]. There are, however, a few exceptions.

Positional Embeddings of Corners are the default choice for SAM-based bounding box approaches [137, 201] which utilize positional embeddings as defined in Equation 33 of the top-left $b_{\text{top-left}}$ and bottom-right $b_{\text{bottom-right}}$ points of the bounding box $\mathcal{B}$ as additional inputs to the model:

$$\mathrm{PE}(\mathcal{B}) = \mathrm{PE}(b_{\text{top-left}}) + \mathrm{PE}(b_{\text{bottom-right}}) \tag{38}$$

Other explicit bounding box signals seen in literature, but only in isolated cases, are Euclidean maps [274], computing the distance to the bounding box boundary as in Equation 7, and constant values filled within the dilation of the bounding box [1] by using the same identity encoding as in Equation 37.

Polygon Vertices-based Explicit Guidance Signals fall into two categories - they either represent the vertices using the identity encoding from Equation 37 over all polygon vertices $\mathcal{P} = \{p_1, \ldots, p_N\}$ and obtain a binary image [58], or they transform the vertex coordinates into an undirected graph $G(\mathcal{P})$ [20, 30, 50] as in Equation 39:

$$G(\mathcal{P}) = (V, E) = \Big(\mathcal{P}, \{\{p_i, p_{i+1}\} \mid i = 1, \ldots, N-1\} \cup \{\{p_N, p_1\}\}\Big) \tag{39}$$

A.2 OMNIMEDSEG DATASETS LIST

Name	Modality	URL	Name	Modality	URL	Name	Modality	URL
PAPILA	Fundus	Link	LES-AV	Fundus	Link	FIVES	Fundus	Link
Fundus-AVSeg	Fundus	Link	NETRA	Fundus	Link	CHAKSU	Fundus	Link
RETA	Fundus	Link	HVDROPDB	Fundus	Link	TREND	Fundus	Link
STARE	Fundus	Link	CHASEDB	Fundus	Link	HRF	Fundus	Link
ORVS	Fundus	Link	RAVIR	Fundus	Link	CSC	Fundus	Link
Akram	Fundus	Link	ROP	Fundus	Link	OcuTox	Fundus	Link
COph100	Fundus	Link	BUSI	Ultrasound	Link	PSFHS	Ultrasound	Link
HC18	Ultrasound	Link	DDTI	Ultrasound	Link	MMOTU	Ultrasound	Link
MuscleUS	Ultrasound	Link	TN3k	Ultrasound	Link	TG3k	Ultrasound	Link
100+ US Images	Ultrasound	Link	FASS	Ultrasound	Link	BUS-BRA	Ultrasound	Link
BUS-UCLM	Ultrasound	Link	OASBUD	Ultrasound	Link	CVC-ClinicDB	Endoscopy	Link
EAD	Endoscopy	Link	Fetoscopy Placenta Data	Endoscopy	Link	FetReg	Endoscopy	Link
Kvasir-Capsule-Seg	Endoscopy	Link	Kvasir-Seg	Endoscopy	Link	PolypDB	Endoscopy	Link
PolypGen	Endoscopy	Link	EndoVis15	Endoscopy	Link	HyperKvasir	Endoscopy	Link
Laryngeal Endoscopic	Endoscopy	Link	BAGLS	Endoscopy	Link	ICF	OCT	Link
AMD-SD	OCT	Link	OIMHS	OCT	Link	AIDK	OCT	Link
OCT Lesion	OCT	Link	OCTID	OCT	Link	SLiMIA	Microscopy	Link
PCMMD	Microscopy	Link	BBBC010	Microscopy	Link	BriFiSeg	Microscopy	Link
WBC	Microscopy	Link	YeaZ	Microscopy	Link	OCCISC	Microscopy	Link
BBBC038	Microscopy	Link	ssTEM	Microscopy	Link	BBBC041Seg	Microscopy	Link
CCAGT	Microscopy	Link	VICAR	Microscopy	Link	EMDS-6	Microscopy	Link
FNC	Microscopy	Link	PTX-498	X-Ray	Link	BTXRD	X-Ray	Link
Hipbone	X-Ray	Link	PanDental	X-Ray	Link	ARCADE	X-Ray	Link
COVID-19-CXR	X-Ray	Link	TeethSeg	X-Ray	Link	CXRAY Jäger	X-Ray	Link
STS-Tooth	X-Ray	Link	Skin Hair	Dermoscopy	Link	ISIC16	Dermoscopy	Link
HAM 10000	Dermoscopy	Link	Nevus	Dermoscopy	Link	ISIC17	Dermoscopy	Link
ISIC18	Dermoscopy	Link	Fusc2021	Dermoscopy	Link	PDDCA	CT	Link
VESSEL12	CT	Link	Abdomen1k	CT	Link	Adrenal-ACC-Ki67	CT	Link
COVID-19 Lung Lesion	CT	Link	COVID-19 CT LIS	CT	Link	NSCLC Radiogenomics	CT	Link
NSCLC Radiomics	CT	Link	MSD Lung Tumor	CT	Link	HCC-TACE-Seg	CT	Link
KiTS23	CT	Link	KiTS19	CT	Link	Lnq2023	CT	Link
MSD Colon	CT	Link	MSD Liver	CT	Link	MSD Pancreas	CT	Link
MSD Spleen	CT	Link	MSD Hepatic Vessels	CT	Link	PleThora	CT	Link
TotalSegmentator	CT	Link	WORD	CT	Link	CHAOS-CT	CT	Link
Flare-CT-2021	CT	Link	Flare-CT-2022	CT	Link	LUNA	CT	Link
Verse2019	CT	Link	Verse2020	CT	Link	ATM	CT	Link
CTPelvic1K	CT	Link	Pancreas CT	CT	Link	CURVAS	CT	Link
ULS	CT	Link	CRML-CT	CT	Link	AeroPath	CT	Link
Couinaud	CT	Link	DAP Atlas	CT	Link	FUMPE	CT	Link
RibSeg V2	CT	Link	TopCow	CT	Link	SpineMets	CT	Link
CT Lymph Nodes	CT	Link	Mediastinal-Str	CT	Link	Pediatric CT Seg	CT	Link
Prostate Edge Cases	CT	Link	COVID-19-20	CT	Link	AutoPET 2024 CT	CT	Link
STACOM-SLAWT	CT	Link	LCTSC	CT	Link	Pancreatic-CT-CBCT-SEG	CT	Link
STS-Tooth	CT	Link	WAW-TACE	CT	Link	PENGWIN	CT	Link
ACDC	MRI	Link	CC-Tumor-Heterogeneity	MRI	Link	CHAOS-MRI	MRI	Link
crossMODA	MRI	Link	ISLES 2022	MRI	Link	MSD-Heart	MRI	Link
MSD-Prostate	MRI	Link	NCI-ISBI	MRI	Link	PI-CAI	MRI	Link
PROMISE12	MRI	Link	Qin-Prostate-Repeatability	MRI	Link	Spine	MRI	Link
SCD	MRI	Link	CDEMRIS	MRI	Link	SegThy	MRI	Link
AtriaSeg 2018	MRI	Link	HVSMR 2.0	MRI	Link	Soft Tissue Sarcoma	MRI	Link
Spider	MRI	Link	UPENN-GBM	MRI	Link	REMIND	MRI	Link
TotalSegmentator-MRI	MRI	Link	HNTSMRG	MRI	Link	PanSegData	MRI	Link
IBD	MRI	Link	Hypo Subfields	MRI	Link	TOM500	MRI	Link
Pituitary Tumor	MRI	Link	RESECT	MRI	Link	GBM-Reservoir	MRI	Link
MedSeg Ventricles	MRI	Link	Prostate-MRI	MRI	Link	BrainPTM	MRI	Link
			STACOM-SLAWT	MRI	Link			

Table 20: List of all datasets included in *OmniMedSeg*. Detailed information regarding licenses, attribution, citation, and dataset descriptions is available at the respective dataset link.

B

AUTHORED PUBLICATIONS

This doctoral research resulted in the following publications related to the thesis (coarsely sorted by relevance, although such order is not easy to define).

* indicates that Zdravko Marinov is an equal first co-author.

1. Zdravko Marinov*, Paul F. Jäger*, Jan Egger, Jens Kleesiek, Rainer Stiefelhagen. **Deep Interactive Segmentation of Medical Images: A Systematic Review and Taxonomy.** *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, August 2024.

2. Zdravko Marinov, Rainer Stiefelhagen, Jens Kleesiek. **Guiding the Guidance: A Comparative Analysis of User Guidance Signals for Interactive Segmentation of Volumetric Images.** *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, October 2023.

3. Matthias Hadlich*, Zdravko Marinov*, Moon Kim, Enrico Nasca, Jens Kleesiek, Rainer Stiefelhagen. **Sliding Window FastEdit: A Framework for Lesion Annotation in Whole-body PET Images.** *International Symposium on Biomedical Imaging (ISBI)*, May 2024.

4. Zdravko Marinov, Moon Kim, Jens Kleesiek, Rainer Stiefelhagen. **Rethinking Annotator Simulation: Realistic Evaluation of Whole-Body PET Lesion Interactive Segmentation Methods.** *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) Workshops*, October 2024.

5. Matthias Hadlich*, Zdravko Marinov*, Rainer Stiefelhagen. **AutoPET Challenge 2023: Sliding Window-based Optimization of U-Net.** *arXiv preprint arXiv:2309.12114*, September 2023, 🏆 **Top 3 position in the autoPET II challenge.**

6. Zdravko Marinov, Alexander Jaus, Jens Kleesiek, Rainer Stiefelhagen. **Filters, Thresholds, and Geodesic Distances for Scribble-based Interactive Segmentation of Medical Images.** *Medical Image Segmentation Foundation Models. CVPR 2024 Challenge: Segment Anything in Medical Images on Laptop*, June 2024, 🏆 **1st place in the Efficient MedSAMs challenge: Task 2.**

7. Zdravko Marinov, Alexander Jaus, Jens Kleesiek, Rainer Stiefelhagen. **Taking a Step Back: Revisiting Classical Approaches for Efficient Interactive Segmentation of Medical Images.** *Medical Image Segmentation Foundation Models. CVPR 2024 Challenge: Segment*

Anything in Medical Images on Laptop, June 2024, 🏆 **Top 6 in the Efficient MedSAMs challenge: Task 1.**

8. Verena Jasmin Hallitschke, Tobias Schlumberger, Philipp Kataliakos, Zdravko Marinov, Moon Kim, Lars Heiliger, Constantin Seibold, Jens Kleesiek, Rainer Stiefelhagen. **Multimodal Interactive Lung Lesion Segmentation: A Framework for Annotating PET/CT Images Based on Physiological and Anatomical Cues.** *International Symposium on Biomedical Imaging (ISBI)*, April 2023, **oral**.

The following scientific publications were co-authored by Zdravko Marinov during his PhD research, but are out of the scope of this thesis (chronological order).

1. Zdravko Marinov, Jens Kleesiek, Rainer Stiefelhagen. **Enforcing Anatomical Symmetry with Euclidean Distance Transforms for Low-Field MRI Bilateral Structure Segmentation.** *Low Field Pediatric Brain Magnetic Resonance Image Segmentation and Quality Assurance: LISA 2025, Held in Conjunction with MICCAI 2025*, (to appear).

2. Zdravko Marinov, Jens Kleesiek, Rainer Stiefelhagen. **Not All Ensembles Are Equal: A Short Study on Moderate to Severe Traumatic Brain Injury Segmentation Methods.** *Automated Identification of Moderate-to-Severe Traumatic Brain Injury Lesions: AIMS-TBI 2025, Held in Conjunction with MICCAI 2025*, (to appear).

3. Zdravko Marinov, Jens Kleesiek, Rainer Stiefelhagen. **U-Net Adapter-based Universal Ultrasound Segmentation and Classification.** *Universal Ultrasound Image Challenge: Multi-Organ Classification and Segmentation: UUSIC 2025, Held in Conjunction with MICCAI 2025*, (to appear), 🏆 **Top 10 position in the UUSIC challenge.**

4. Jun Ma et al. **Efficient MedSAMs: Segment Anything in Medical Images on Laptop.**, (under review).

5. Alexander Jaus, Zdravko Marinov, Jiale Wei, Simon Reiß, Rainer Stiefelhagen. **Applications of High-Performance Computing for Medical Image Analysis.** *Transactions of the High Performance Computing Center, Stuttgart (HLRS) 2025*, (to appear).

6. Simon Reiß, Zdravko Marinov, Alexander Jaus, Constantin Seibold, M Saquib Sarfraz, Erik Rodner, Rainer Stiefelhagen. **Is Visual in-Context Learning for Compositional Medical Tasks within Reach?** *International Conference on Computer Vision (ICCV)*, October 2025.

7. Lena Heinemann, Alexander Jaus, Zdravko Marinov, Moon Kim, Maria Francesca Spadea, Jens Kleesiek, Rainer Stiefelhagen. **LIMIS: Towards Language-Based Interactive Medical Image Segmentation.** *International Symposium on Biomedical Imaging (ISBI)*, April 2025.

8. Alexander Jaus, Constantin Marc Seibold, Simon Reiß, Zdravko Marinov, Keyi Li, Zeling Ye, Stefan Krieg, Jens Kleesiek, Rainer Stiefelhagen. **Every Component Counts: Rethinking the Measure of Success for Medical Semantic Segmentation in Multi-Instance Segmentation Tasks.** *AAAI Conference on Artificial Intelligence*, April 2025.

9. Jianning Li et al. **MedShapeNet – A Large-scale Dataset of 3D Medical Shapes for Computer Vision.** *Biomedical Engineering/Biomedizinische Technik*, February 2025.

10. Zdravko Marinov, Simon Reiß, David Kersting, Jens Kleesiek, Rainer Stiefelhagen. **Mirror U-Net: Marrying Multimodal Fission with Multi-task Learning for Semantic Segmentation in Medical Imaging.** *International Conference on Computer Vision (ICCV) Workshops*, October 2023, **oral**.

11. Sergios Gatidis et al. **Results from the autoPET Challenge on Fully Automated Lesion Segmentation in Oncologic PET/CT Imaging.** *Nature Machine Intelligence*, November 2024.

12. Valentin Khan-Blouki, Franziska Seiz, Nicolas Walter, Alexander Jaus, Zdravko Marinov, Gijs Luijten, Jan Egger, Constantin Marc Seibold, Dirk Solte, Jens Kleesiek, Rainer Stiefelhagen. **FootCapture: Towards an AR-based System for 3D Foot Object Acquisition through Photogrammetry.** *Medical Imaging with Deep Learning (MIDL)*, July 2024.

13. Calvin Tanama, Kunyu Peng, Zdravko Marinov, Rainer Stiefelhagen, Alina Roitberg. **Quantized Distillation: Optimizing Driver Activity Recognition Models for Resource-Constrained Environments.** *International Conference on Intelligent Robots and Systems (IROS)*, October 2023, **oral**.

14. Zdravko Marinov, David Schneider, Alina Roitberg, Rainer Stiefelhagen. **ModSelect: Automatic Modality Selection for Synthetic-to-Real Domain Generalization.** *European Conference on Computer Vision (ECCV) Workshops*, October 2022, **oral**.

15. Zdravko Marinov*, David Schneider*, Alina Roitberg*, Rainer Stiefelhagen. **Multimodal Generation of Novel Action Appearances for Synthetic-to-Real Recognition of Activities of Daily Living.** *International Conference on Intelligent Robots and Systems (IROS)*, October 2022, **oral**.

16. Lars Heiliger*, Zdravko Marinov*, Max Hasin, André Ferreira, Jana Fragemann, Kelsey Pomykala, Jacob Murray, David Kersting, Victor Alves, Rainer Stiefelhagen, Jan Egger, Jens Kleesiek. **AutoPET Challenge: Combining nn-Unet with Swin UNETR Augmented by Maximum Intensity Projection Classifier.** *arXiv preprint arXiv:2209.01112*, September 2022, 🏆 **Top 5 position in the autoPET challenge.**

17. Alina Roitberg, Kunyu Peng, Zdravko Marinov, Constantin Seibold, David Schneider, Rainer Stiefelhagen. **A Comparative Analysis of Decision-Level Fusion for Multimodal Driver Behaviour Understanding.** *Intelligent Vehicles Symposium (IV)*, June 2022.

C

DISCLOSURE OF THE USE OF GENERATIVE AI AND AI-ASSISTED TOOLS IN THE WRITING PROCESS

Following the "Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG" (Link) from September 2023, the author (Zdravko Marinov) utilized ChatGPT (version 4, 4.5, and 5) to enhance the language quality of this work and to support typesetting tasks (e.g., formatting tables).

BIBLIOGRAPHY

[1] Martin Rajchl et al. DeepCut: Object Segmentation from Bounding Box Annotations Using Convolutional Neural Networks. In: *IEEE Transactions on Medical Imaging* 36.2 (2016), pp. 674–683.

[2] Mario Amrehn et al. UI-net: Interactive artificial neural networks for iterative image segmentation based on a user model. In: *Eurographics Workshop on Visual Computing for Biology and Medicine* (2017), pp. 143–147.

[3] Jinquan Sun, Yinghuan Shi, Yang Gao, and Dinggang Shen. A point says a lot: an interactive segmentation method for MR prostate via one-point labeling. In: *International Workshop on Machine Learning in Medical Imaging* (2017), pp. 220–228.

[4] Yigit B Can et al. Learning to segment medical images with scribble-supervision alone. In: *International Workshop on Deep Learning in Medical Image Analysis* (2018), pp. 236–244.

[5] Guotai Wang et al. DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.7 (2018), pp. 1559–1572.

[6] Guotai Wang et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. In: *IEEE Transactions on Medical Imaging* 37.7 (2018), pp. 1562–1573.

[7] Gustav Bredell, Christine Tanner, and Ender Konukoglu. Iterative interaction training for segmentation editing networks. In: *International Workshop on Machine Learning in Medical Imaging* (2018), pp. 363–370.

[8] Ashis Kumar Dhara et al. Segmentation of post-operative glioblastoma in MRI by U-Net with patient-specific interactive refinement. In: *International MICCAI Brainlesion Workshop* (2019), pp. 115–122.

[9] Youbao Tang, Adam P Harrison, Mohammadhadi Bagheri, Jing Xiao, and Ronald M Summers. Semi-automatic RECIST labeling on CT scans with cascaded convolutional neural networks. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), pp. 405–413.

[10] Tomas Sakinis et al. Interactive segmentation of medical images through fully convolutional neural networks. In: *arXiv:1903.08205v1* (2019).

[11] Bowei Zhou, Li Chen, and Zhao Wang. Interactive deep editing framework for medical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), pp. 329–337.

[12] Shadab Khan, Ahmed H Shahin, Javier Villafruela, Jianbing Shen, and Ling Shao. Extreme points derived confidence map as a cue for class-agnostic interactive segmentation using deep neural network. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2019), pp. 66–73.

[13] Wenhui Lei, Huan Wang, Ran Gu, Shichuan Zhang, Shaoting Zhang, and Guotai Wang. DeepIGeoS-V2: deep interactive segmentation of multiple organs from head and neck images with lightweight CNNs. In: *LABELS Workshop MICCAI* (2019), pp. 61–69.

[14] Guilherme Aresta et al. iW-Net: An automatic and minimalistic interactive lung nodule segmentation deep network. In: *Scientific Reports* 9.1 (2019), pp. 1–9.

[15] Holger Roth et al. Weakly supervised segmentation from extreme points. In: *LABELS Workshop MICCAI* (2019), pp. 42–50.

[16] Lorenzo Cerrone, Alexander Zeilmann, and Fred A Hamprecht. End-to-end learned random walker for seeded image segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 12559–12568.

[17] Haiyan Zheng, Yufei Chen, Xiaodong Yue, and Chao Ma. Deep interactive segmentation of uncertain regions with shadowed sets. In: *Third International Symposium on Image Computing and Digital Medicine* (2019), pp. 244–248.

[18] Chun-Hung Chao et al. Radiotherapy Target Contouring with Convolutional Gated Graph Neural Network. In: *arXiv:1904.03086v1* (2019).

[19] Martin Längkvist, Jonas Widell, Per Thunberg, Amy Loutfi, and Mats Lidén. Interactive user interface based on Convolutional Auto-encoders for annotating CT-scans. In: *arXiv:1904.11701v1* (2019).

[20] Xiaosong Wang, Ling Zhang, Holger Roth, Daguang Xu, and Ziyue Xu. Interactive 3D segmentation editing and refinement via gated graph neural networks. In: *International Workshop on Graph Learning in Medical Imaging* (2019), pp. 9–17.

[21] TGW Boers et al. Interactive 3D U-net for the segmentation of the pancreas in computed tomography scans. In: *Physics in Medicine & Biology* 65.6 (2020, Art. no. 065002).

[22] Guotai Wang, Michael Aertsen, Jan Deprest, Sébastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of MRI slices. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), pp. 279–288.

[23] Xuan Liao et al. Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9394–9402.

[24] Ashwin Raju et al. User-guided domain adaptation for rapid annotation from user interactions: a study on pathological liver segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), pp. 457–467.

[25] Chaofan Ma et al. Boundary-aware supervoxel-level iteratively refined interactive 3D image segmentation with multi-agent reinforcement learning. In: *IEEE Transactions on Medical Imaging* 40.10 (2020), pp. 2563–2574.

[26] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. NuClick: a deep learning framework for interactive segmentation of microscopic images. In: *Medical Image Analysis* 65 (2020, Art. no. 101771).

[27] Titinunt Kitrungrotsakul, Iwamoto Yutaro, Lanfen Lin, Ruofeng Tong, Jingsong Li, and Yen-Wei Chen. Interactive deep refinement network for medical image segmentation. In: *arXiv:2006.15320v1* (2020).

[28] Antonio Pepe et al. IRIS: Interactive real-time feedback image segmentation with deep learning. In: *SPIE Medical Imaging: Biomedical Applications in Molecular, Structural, and Functional Imaging* 11317 (2020), pp. 181–186.

[29] Weifeng Hu et al. Error Attention Interactive Segmentation of Medical Image Through Matting and Fusion. In: *International Workshop on Machine Learning in Medical Imaging* (2020), pp. 11–20.

[30] Zhiqiang Tian et al. Graph-convolutional-network-based interactive prostate segmentation in MR images. In: *Medical Physics* 47.9 (2020), pp. 4164–4176.

[31] Chun-Hung Chao, Hsien-Tzu Cheng, Tsung-Ying Ho, Le Lu, and Min Sun. Interactive radiotherapy target delineation with 3d-fused context propagation. In: *arXiv:2012.06873v1* (2020).

[32] Youbao Tang, Ke Yan, Jing Xiao, and Ronald M Summers. One click lesion RECIST measurement and segmentation on CT scans. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), pp. 573–583.

[33] Hikari Jinbo, Titinunt Kitrungrotsaku, Yutaro Iwamoto, Lanfen Lin, Hongjie Hu, and Yen-Wei Chen. Development of an Interactive Semantic Medical Image Segmentation System. In: *IEEE Global Conference on Consumer Electronics* (2020), pp. 678–681.

[34] Kibrom Berihu Girum, Gilles Créhange, Raabid Hussain, and Alain Lalande. Fast interactive medical image segmentation with weakly supervised deep learning method. In: *International Journal of Computer Assisted Radiology and Surgery* 15 (2020), pp. 1437–1444.

[35] David Joon Ho et al. Deep interactive learning: an efficient labeling approach for deep learning-based osteosarcoma treatment response assessment. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2020), pp. 540–549.

[36] Michelle Xiao-Lin Foo et al. Interactive Segmentation for COVID-19 Infection Quantification on Longitudinal CT scans. In: *IEEE Access* 11 (2023), pp. 77596–77607.

[37] Ashish Menon, Piyush Singh, PK Vinod, and CV Jawahar. Interactive Learning for Assisting Whole Slide Image Annotation. In: *Asian Conference on Pattern Recognition* (2021), pp. 504–517.

[38] Xiangde Luo et al. MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning. In: *Medical Image Analysis* 72 (2021, Art. no. 102102).

[39] Ruiwei Feng et al. Interactive few-shot learning: Limited supervision, better medical image segmentation. In: *IEEE Transactions on Medical Imaging* 40.10 (2021), pp. 2575–2588.

[40] Holger R Roth, Dong Yang, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Going to extremes: weakly supervised medical image segmentation. In: *Machine Learning and Knowledge Extraction* 3.2 (2021), pp. 507–524.

[41] Bhavani Sambaturu, Ashutosh Gupta, CV Jawahar, and Chetan Arora. Efficient and Generic Interactive Segmentation Framework to Correct Mispredictions during Clinical Evaluation of Medical Images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2021), pp. 625–635.

[42] Tianfei Zhou, Liulei Li, Gustav Bredell, Jianwu Li, and Ender Konukoglu. Quality-aware memory network for interactive volumetric image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2021), pp. 560–570.

[43] Helena Williams et al. Interactive segmentation via deep learning and B-spline explicit active surfaces. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2021), pp. 315–325.

[44] Xiaokang Li et al. WDTIseg: One-stage interactive segmentation for breast ultrasound image using weighted distance transform and shape-aware compound loss. In: *Applied Sciences* 11.14 (2021, Art. no. 6279).

[45] Wenhao Li et al. Interactive medical image segmentation with self-adaptive confidence calibration. In: *Frontiers of Information Technology & Electronic Engineering* 24.9 (2023), pp. 1332–1348.

[46] Jingjing Deng and Xianghua Xie. 3D Interactive Segmentation With Semi-Implicit Representation and Active Learning. In: *IEEE Transactions on Image Processing* 30 (2021), pp. 9402–9417.

[47] Jian Zhang et al. Interactive medical image segmentation via a point-based interaction. In: *Artificial Intelligence in Medicine* 111 (2021, Art. no. 101998).

[48] Ervine Zheng, Qi Yu, Rui Li, Pengcheng Shi, and Anne Haake. A continual learning framework for uncertainty-aware interactive image segmentation. In: *AAAI Conference on Artificial Intelligence* (2021), pp. 6030–6038.

[49] Jian-Wei Zhang et al. DINs: deep interactive networks for neurofibroma segmentation in neurofibromatosis type 1 on whole-body mri. In: *IEEE Journal of Biomedical and Health Informatics* 26.2 (2021), pp. 786–797.

[50] Zhiqiang Tian et al. Interactive prostate MR image segmentation based on ConvLSTMs and GGNN. In: *Neurocomputing* 438 (2021), pp. 84–93.

[51] Dalei Jiang et al. Residual refinement for interactive skin lesion segmentation. In: *Journal of Biomedical Semantics* 12.1 (2021, Art. no. 22).

[52] Yunkun Bai, Guangmin Sun, Yu Li, Le Shen, and Li Zhang. Progressive medical image annotation with convolutional neural network-based interactive segmentation method. In: *SPIE Medical Imaging: Image Processing* 11596 (2021), pp. 732–742.

[53] Sungduk Cho, Hyungjoon Jang, Jing Wei Tan, and Won-Ki Jeong. DeepScribble: interactive pathology image segmentation using deep neural networks with scribbles. In: *IEEE International Symposium on Biomedical Imaging* (2021), pp. 761–765.

[54] Titinunt Kitrungrotsakul et al. Attention-RefNet: Interactive attention refinement network for infected area segmentation of COVID-19. In: *IEEE Journal of Biomedical and Health Informatics* 25.7 (2021), pp. 2363–2373.

[55] Rajshree Daulatabad, Roberto Vega, Jacob L Jaremko, Jeevesh Kapur, Abhilash R Hareendranathan, and Kumaradeven Punithakumar. Integrating User-Input into Deep Convolutional Neural Networks for Thyroid Nodule Segmentation. In: *International Conference of the IEEE Engineering in Medicine & Biology Society* (2021), pp. 2637–2640.

[56] Xuan Huy Manh et al. Interactive Z-line segmentation tool for Upper Gastrointestinal Endoscopy Images using Binary Partition Tree and U-Net. In: *RIVF International Conference on Computing and Communication Technologies* (2021), pp. 1–6.

[57] Michael J Trimpl, Djamal Boukerroui, Eleanor PJ Stride, Katherine A Vallis, and Mark J Gooding. Interactive contouring through contextual deep learning. In: *Medical Physics* 48.6 (2021), pp. 2951–2959.

[58] Yuqi Fang, Delong Zhu, Niyun Zhou, Li Liu, and Jianhua Yao. PiPo-Net: A Semi-automatic and Polygon-based Annotation Method for Pathological Images. In: *IEEE/RSJ International Conference on Intelligent Robots and Systems* (2021), pp. 2978–2984.

[59] Mostafa Jahanifar, Neda Zamani Tajeddin, Navid Alemi Koohbanani, and Nasir M Rajpoot. Robust interactive semantic segmentation of pathology images with minimal user input. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 674–683.

[60] Lei Sun, Zhiqiang Tian, Zhang Chen, Wenrui Luo, and Shaoyi Du. An efficient interactive segmentation framework for medical images without pre-training. In: *Medical Physics* 50.4 (2023), pp. 2239–2248.

[61] Maysam Shahedi, James D Dormer, Martin Halicek, and Baowei Fei. The effect of image annotation with minimal manual interaction for semiautomatic prostate segmentation in CT images using fully convolutional neural networks. In: *Medical Physics* 49.2 (2022), pp. 1153–1160.

[62] Alessia Atzeni et al. Deep active learning for suggestive segmentation of biomedical image stacks via optimisation of Dice scores and traced boundary length. In: *Medical Image Analysis* 81 (2022, Art. no. 102549).

[63] Lei Bi, Michael Fulham, and Jinman Kim. Hyper-fusion network for semi-automatic segmentation of skin lesions. In: *Medical Image Analysis* 76 (2022, Art. no. 102334).

[64] Qin Liu, Zhenlin Xu, Yining Jiao, and Marc Niethammer. iSegFormer: Interactive Segmentation via Transformers with Application to 3D Knee MR Images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), pp. 464–474.

[65] Muhammad Asad, Lucas Fidon, and Tom Vercauteren. ECONet: Efficient convolutional online likelihood network for scribble-based interactive segmentation. In: *International Conference on Medical Imaging with Deep Learning* (2022), pp. 35–47.

[66] Karol Gotkowski, Camila Gonzalez, Isabel Kaltenborn, Ricarda Fischbach, Andreas Bucher, and Anirban Mukhopadhyay. i3Deep: Efficient 3D interactive segmentation with the nnU-Net. In: *International Conference on Medical Imaging with Deep Learning* (2022), pp. 441–456.

[67] Andres Diaz-Pinto et al. DeepEdit: Deep Editable Learning for Interactive Segmentation of 3D Medical Images. In: *DALI Workshop MICCAI* (2022), pp. 11–21.

[68] Wentao Liu, Chaofan Ma, Yuhuan Yang, Weidi Xie, and Ya Zhang. Transforming the Interactive Segmentation for Medical Imaging. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), pp. 704–713.

[69] Luyue Shi, Xuanye Zhang, Yunbi Liu, and Xiaoguang Han. A Hybrid Propagation Network for Interactive Volumetric Image Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2022), pp. 673–682.

[70] Mingrui Zhuang et al. AnatomySketch: An Extensible Open-Source Software Platform for Medical Image Analysis Algorithm Development. In: *Journal of Digital Imaging* 35.6 (2022), pp. 1623–1633.

[71] Gaëtan Galisot, Jean-Yves Ramel, Thierry Brouard, Elodie Chaillou, and Barthélémy Serres. Visual and structural feature combination in an interactive machine learning system for medical image segmentation. In: *Machine Learning with Applications* 8 (2022, Art. no. 100294).

[72] Zheng Lin, Zhao Zhang, Ling-Hao Han, and Shao-Ping Lu. Multi-Mode Interactive Image Segmentation. In: *ACM International Conference on Multimedia* (2022), pp. 905–914.

[73] Ragavie Pirabaharan and Naimul Khan. Interactive segmentation using U-Net with weight map and dynamic user interactions. In: *International Conference of the IEEE Engineering in Medicine & Biology Society* (2022), pp. 4754–4757.

[74] Ivan Mikhailov, Benoit Chauveau, Nicolas Bourdel, and Adrien Bartoli. A Deep Learning-Based Interactive Medical Image Segmentation Framework. In: *AMAI Workshop MICCAI* (2022), pp. 98–107.

[75] Ragavie Pirabaharan and Naimul Khan. Improving Interactive Segmentation using a Novel Weighted Loss Function with an Adaptive Click Size and Two-Stream Fusion. In: *IEEE International Conference on Multimedia Big Data* (2022), pp. 7–12.

[76] Xuan Chen et al. Balancing regional and global information: An interactive segmentation framework for ultrasound breast lesion. In: *Biomedical Signal Processing and Control* 77 (2022, Art. no. 103723).

[77] Shi Liang et al. Deep SED-Net with interactive learning for multiple testicular cell types segmentation and cell composition analysis in mouse seminiferous tubules. In: *Cytometry Part A* 101.8 (2022), pp. 658–674.

[78] Mingeon Ju, Moonhyun Lee, Jaeyoung Lee, Jaewoo Yang, Seunghan Yoon, and Younghoon Kim. All You Need Is a Few Dots to Label CT Images for Organ Segmentation. In: *Applied Sciences* 12.3 (2022, Art. no. 1328).

[79] Wenao Ma, Shuang Zheng, Lei Zhang, Huimao Zhang, and Qi Dou. Rapid model transfer for medical image segmentation via iterative human-in-the-loop update: from labelled public to unlabelled clinical datasets for multi-organ segmentation in CT. In: *IEEE International Symposium on Biomedical Imaging* (2022), pp. 1–5.

[80] Ti Bai et al. A Proof-of-Concept Study of Artificial Intelligence–assisted Contour Editing. In: *Radiology: Artificial Intelligence* 4.5 (2022, Art. no. e210214).

[81] Tianfei Zhou, Liulei Li, Gustav Bredell, Jianwu Li, Jan Unkelbach, and Ender Konukoglu. Volumetric memory network for interactive medical image segmentation. In: *Medical Image Analysis* 83 (2023, Art. no. 102599).

[82] Verena Jasmin Hallitschke et al. Multimodal Interactive Lung Lesion Segmentation: A Framework for Annotating PET/CT Images based on Physiological and Anatomical Cues. In: *IEEE International Symposium on Biomedical Imaging* (2023), pp. 1–5.

[83] Qin Liu et al. Exploring Cycle Consistency Learning in Interactive Volume Segmentation. In: *arXiv:2303.06493v2* (2023).

[84] Aldimir Bruzadin, Maurílio Boaventura, Marilaine Colnago, Rogério Galante Negri, and Wallace Casaca. Learning label diffusion maps for semi-automatic segmentation of lung CT images with COVID-19. In: *Neurocomputing* 522 (2023), pp. 24–38.

[85] Muhammad Asad et al. Adaptive Multi-scale Online Likelihood Network for AI-assisted Interactive Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023).

[86] Ahmed H Shahin, Yan Zhuang, and Noha El-Zehiry. From Sparse to Precise: A Practical Editing Approach for Intracardiac Echocardiography Segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023).

[87] Mingrui Zhuang et al. Efficient contour-based annotation by iterative deep learning for organ segmentation from volumetric medical images. In: *International Journal of Computer Assisted Radiology and Surgery* 18.2 (2022), pp. 379–394.

[88] David Joon Ho et al. Deep Interactive Learning-based ovarian cancer segmentation of H&E-stained whole slide images to study morphological patterns of BRCA mutation. In: *Journal of Pathology Informatics* 14 (2023, Art. no. 100160).

[89] Zixiang Wei, Jintao Ren, Stine Sofia Korreman, and Jasper Nijkamp. Towards interactive deep-learning for tumour segmentation in head and neck cancer radiotherapy. In: *Physics and Imaging in Radiation Oncology* 25 (2023, Art. no. 100408), p. 100408.

[90] Mingrui Zhuang, Zhonghua Chen, Yuxin Yang, Lauri Kettunen, and Hongkai Wang. Annotation-efficient training of medical image segmentation network based on scribble guidance in difficult areas. In: *International Journal of Computer Assisted Radiology and Surgery* 19.1 (2023), pp. 87–96.

[91] Zdravko Marinov, Rainer Stiefelhagen, and Jens Kleesiek. Guiding the Guidance: A Comparative Analysis of User Guidance Signals for Interactive Segmentation of Volumetric Images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2023), pp. 637–647.

[92] Chongyu Qu et al. AbdomenAtlas-8K: Annotating 8,000 CT Volumes for Multi-Organ Segmentation in Three Weeks. In: *Advances in Neural Information Processing Systems* (2023), pp. 36620–36636.

[93] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for Medical Image Analysis: an experimental study. In: *Medical Image Analysis* 89 (2023, Art. no. 102918).

[94] Ruining Deng et al. Segment anything model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging. In: *International Conference on Medical Imaging with Deep Learning, Short Paper Track* (2023).

[95] S Mohapatra, A Gosai, and G Schlaug. SAM vs BET: A comparative study for brain extraction and segmentation of magnetic resonance images using deep learning. In: *arXiv:2304.04738v3* (2023).

[96] Florian Putz et al. The Segment Anything foundation model achieves favorable brain tumor autosegmentation accuracy on MRI to support radiotherapy treatment planning. In: *arXiv:2304.07875v1* (2023).

[97] Chuanfei Hu and Xinde Li. When SAM meets medical images: An investigation of segment anything model (SAM) on multi-phase liver tumor segmentation. In: *arXiv:2304.08506v6* (2023).

[98] Tianrun Chen et al. SAM Fails to Segment Anything?–SAM-Adapter: Adapting SAM in Underperformed Scenes: Camouflage, Shadow, and More. In: *arXiv:2304.09148v3* (2023).

[99] Junde Wu et al. Medical SAM adapter: Adapting segment anything model for medical image segmentation. In: *arXiv:2304.12620v7* (2023).

[100] Zhongxi Qiu, Yan Hu, Heng Li, and Jiang Liu. Learnable ophthalmology SAM. In: *arXiv:2304.13425v1* (2023).

[101] Sheng He, Rina Bao, Jingpeng Li, P Ellen Grant, and Yangming Ou. Accuracy of segment-anything model (SAM) in medical image segmentation tasks. In: *arXiv:2304.09324v3* (2023).

[102] Peilun Shi, Jianing Qiu, Sai Mu Dalike Abaxi, Hao Wei, Frank P-W Lo, and Wu Yuan. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. In: *Diagnostics* 13.11 (2023, Art. no. 1947).

[103] Bin Wang, Armstrong Aboah, Zheyuan Zhang, and Ulas Bagci. GazeSAM: What you see is what you segment. In: *arXiv:2304.13844v1* (2023).

[104] Mingzhe Hu, Yuheng Li, and Xiaofeng Yang. SkinSAM: Empowering skin cancer segmentation with segment anything model. In: *arXiv:2304.13973v1* (2023).

[105] An Wang, Mobarakol Islam, Mengya Xu, Yang Zhang, and Hongliang Ren. SAM meets robotic surgery: An empirical study in robustness perspective. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention Workshops* (2023), pp. 234–244.

[106] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. SAM on medical images: A comprehensive study on three prompt modes. In: *arXiv:2305.00035v1* (2023).

[107] Christian Mattjie et al. Exploring the zero-shot capabilities of the segment anything model (SAM) in 2D medical imaging: A comprehensive evaluation and practical guideline. In: *arXiv:2305.00109v2* (2023).

[108] Yuheng Li, Mingzhe Hu, and Xiaofeng Yang. Polyp-SAM: Transfer SAM for polyp segmentation. In: *arXiv:2305.00293v1* (2023).

[109] Junde Wu, Jiayuan Zhu, Yuanpei Liu, Yueming Jin, and Min Xu. One-Prompt to Segment All Medical Images. In: *arXiv:2305.10300v3* (2023).

[110] Mingzhe Hu, Yuheng Li, and Xiaofeng Yang. BreastSAM: A study of segment anything model for breast tumor detection in ultrasound images. In: *arXiv:2305.12447v1* (2023).

[111] Dongjoo Lee, Jeongbin Park, Seungho Cook, seong-jin Yoo, Daeseung Lee, and Hongyoon Choi. IAMSAM: Image-based Analysis of Molecular signatures using the Segment-Anything Model. In: *bioRxiv* (2023).

[112] Yifan Gao, Wei Xia, Dingdu Hu, and Xin Gao. DeSAM: Decoupling Segment Anything Model for Generalizable Medical Image Segmentation. In: *arXiv:2306.00499v1* (2023).

[113] Chuyun Shen, Wenhao Li, Ya Zhang, and Xiangfeng Wang. Temporally-Extended Prompts Optimization for SAM in Interactive Medical Image Segmentation. In: *IEEE International Conference on Bioinformatics and Biomedicine* (2023), pp. 3550–3557.

[114] Guochen Ning, Hanyin Liang, Zhongliang Jiang, Hui Zhang, and Hongen Liao. The potential of 'Segment Anything'(SAM) for universal intelligent ultrasound image guidance. In: *Bioscience Trends* 17.3 (2023), pp. 230–233.

[115] Lian Zhang et al. Segment Anything Model (SAM) for Radiation Oncology. In: *arXiv:2306.11730v2* (2023).

[116] Wenhui Lei, Xu Wei, Xiaofan Zhang, Kang Li, and Shaoting Zhang. MedLSAM: Localize and Segment Anything Model for 3D Medical Images. In: *arXiv:2306.14752v3* (2023).

[117] Guoyao Deng et al. SAM-U: Multi-box prompts triggered uncertainty estimation for reliable SAM in medical image. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention Workshops* (2023), pp. 368–377.

[118] Shizhan Gong et al. 3DSAM-adapter: Holistic Adaptation of SAM from 2D to 3D for Promptable Medical Image Segmentation. In: *arXiv:2306.13465v1* (2023).

[119] Yuhao Huang et al. Segment anything model for medical images? In: *Medical Image Analysis* 92 (2024, Art. no. 103061).

[120] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. In: *Nature Communications* 15.1 (2024, Art. no. 654).

[121] Saikat Roy et al. SAM.MD: Zero-shot medical image segmentation capabilities of the segment anything model. In: *International Conference on Medical Imaging with Deep Learning, Short Paper Track* (2023).

[122] Hannes Nickisch, Carsten Rother, Pushmeet Kohli, and Christoph Rhemann. Learning an interactive segmentation system. In: *Indian Conference on Computer Vision, Graphics and Image Processing* (2010), pp. 274–281.

[123] John Canny. A computational approach to edge detection. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6 (1986), pp. 679–698.

[124] Witold Pedrycz. Shadowed sets: representing and processing fuzzy sets. In: *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 28.1 (1998), pp. 103–109.

[125] Daniel Barbosa, Thomas Dietenbeck, Joel Schaerer, Jan D'hooge, Denis Friboulet, and Olivier Bernard. B-spline explicit active surfaces: an efficient framework for real-time 3-D region-based segmentation. In: *IEEE Transactions on Image Processing* 21.1 (2011), pp. 241–251.

[126] Anthony Yezzi Jr, Andy Tsai, and Alan Willsky. A fully global approach to image segmentation via coupled curve evolution equations. In: *Journal of Visual Communication and Image Representation* 13.1 (2002), pp. 195–216.

[127] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015), pp. 234–241.

[128] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[129] Zhi-Hua Zhou. A brief introduction to weakly supervised learning. In: *Nature Science Review* 5.1 (2018), pp. 44–53.

[130] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141.

[131] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in ND images. In: *IEEE/CVF International Conference on Computer Vision* 1 (2001), pp. 105–112.

[132] Philippe Salembier and Luis Garrido. Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval. In: *IEEE Transactions on Image Processing* 9.4 (2000), pp. 561–576.

[133] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (2012), pp. 2274–2282.

[134] Mark Jenkinson et al. BET2: MR-based estimation of brain, skull and scalp surfaces. In: *11th Annual Meeting of the Organization for Human Brain Mapping* (2005, Art. no. 167).

[135] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: *International MICCAI Brainlesion Workshop* (2021), pp. 272–284.

[136] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. In: *Nature Methods* 18.2 (2020), pp. 203–211.

[137] Alexander Kirillov et al. Segment anything. In: *IEEE/CVF International Conference on Computer Vision* (2023), pp. 4015–4026.

[138] Alexey Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale. In: *International Conference on Learning Representations* (2021).

[139] David Moher, Alessandro Liberati, Jennifer Tetzlaff, and Douglas G Altman. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. In: *Bmj* 339 (2009).

[140] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2016), pp. 373–381.

[141] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "GrabCut-interactive foreground extraction using iterated graph cuts. In: *ACM Transactions on Graphics* 23.3 (2004), pp. 309–314.

[142] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2016), pp. 724–732.

[143] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The PASCAL visual object classes (VOC) challenge. In: *International Journal of Computer Vision* 88 (2009), pp. 303–338.

[144] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In: *IEEE/CVF International Conference on Computer Vision* (2011), pp. 991–998.

[145] Kevin McGuinness and Noel E O'connor. A comparative evaluation of interactive segmentation algorithms. In: *Pattern Recognition* 43.2 (2010), pp. 434–444.

[146] Matthias Eisenmann et al. Why is the winner the best? In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 19955–19966.

[147] Sandra G Hart and Lowell E Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: *Advances in Psychology* 52 (1988), pp. 139–183.

[148] John Brooke. SUS-A quick and dirty usability scale. In: *Usability Evaluation in Industry* 189.194 (1996), pp. 4–7.

[149] Andres Diaz-Pinto et al. Monai label: A framework for ai-assisted interactive labeling of 3d medical images. In: *Medical Image Analysis* 95 (2024), p. 103207.

[150] Kenneth A Philbrick et al. RIL-contour: a medical imaging dataset annotation tool for and with deep learning. In: *Journal of Digital Imaging* 32 (2019), pp. 571–581.

[151] Philipp D Lösel et al. Introducing Biomedisa as an open-source online platform for biomedical image segmentation. In: *Nature Communications* 11.1 (2020, Art. no. 5577).

[152] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively Trained Interactive Segmentation. In: *British Machine Vision Conference* (2018).

[153] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. Snakes: Active contour models. In: *International Journal of Computer Vision* 1.4 (1988), pp. 321–331.

[154] Sergios Gatidis et al. Results from the autoPET challenge on fully automated lesion segmentation in oncologic PET/CT imaging. In: *Nature Machine Intelligence* 6.11 (2024), pp. 1396–1405.

[155] Feng Zhao and Xianghua Xie. An overview of interactive medical image segmentation. In: *Annals of the BMVA* 2013.7 (2013), pp. 1–22.

[156] Sílvia Delgado Olabarriaga and Arnold WM Smeulders. Interaction in the segmentation of medical images: A survey. In: *Medical Image Analysis* 5.2 (2001), pp. 127–142.

[157] Hiba Ramadan, Chaymae Lachqar, and Hamid Tairi. A survey of recent interactive image segmentation methods. In: *Computational Visual Media* 6 (2020), pp. 355–384.

[158] Çağrı Kaymak and Ayşegül Uçar. A brief survey and an application of semantic image segmentation for autonomous driving. In: *Handbook of Deep Learning Applications* (2019), pp. 161–200.

[159] Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. In: *Journal of Manufacturing Systems* 31.3 (2019), pp. 759–776.

[160] Geert Litjens et al. A survey on deep learning in Medical Image Analysis. In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[161] Mihalj Bakator and Dragica Radosav. Deep learning and medical diagnosis: A review of literature. In: *Multimodal Technologies and Interaction* 2.3 (2018), p. 47.

[162] Bjoern H Menze et al. The multimodal brain tumor image segmentation benchmark (BRATS). In: *IEEE Transactions on Medical Imaging* 34.10 (2014), pp. 1993–2024.

[163] Michela Antonelli et al. The medical segmentation decathlon. In: *Nature Communications* 13.1 (2022, Art. no. 4128).

[164] Patrick Bilic et al. The liver tumor segmentation benchmark (LiTS). In: *Medical Image Analysis* 84 (2023, Art. no. 102680), p. 102680.

[165] Lena Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. In: *Nature Methods* 21.2 (2024), pp. 195–212.

[166] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected CRFs with gaussian edge potentials. In: *Advances in Neural Information Processing Systems* 24 (2011).

[167] Aaron E Lefohn, Joshua E Cates, and Ross T Whitaker. Interactive, GPU-based level sets for 3D segmentation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2003), pp. 564–572.

[168] Christoph Sommer, Christoph Straehle, Ullrich Koethe, and Fred A Hamprecht. Ilastik: Interactive learning and segmentation toolkit. In: *IEEE International Symposium on Biomedical Imaging* (2011), pp. 230–233.

[169] Paul A Yushkevich, Yang Gao, and Guido Gerig. ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images. In: *International Conference of the IEEE Engineering in Medicine & Biology Society* (2016), pp. 3342–3345.

[170] Annika Reinke et al. Understanding metric-related pitfalls in image analysis validation. In: *Nature Methods* 21.2 (2024), pp. 182–194.

[171] Stanislav Nikolov et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. In: *Journal of Medical Internet Research* 23.7 (2021, Art. no. e26151).

[172] Ruizhe Li and Xin Chen. An efficient interactive multi-label segmentation tool for 2D and 3D medical images using fully connected conditional random field. In: *Computer Methods and Programs in Biomedicine* 213 (2022, Art. no. 106534).

[173] Ivo Wolf et al. The medical imaging interaction toolkit. In: *Medical Image Analysis* 9.6 (2005), pp. 594–604.

[174] Guotai Wang et al. PyMIC: A deep learning toolkit for annotation-efficient medical image segmentation. In: *Computer Methods and Programs in Biomedicine* 231 (2023, Art. no. 107398).

[175] Lluis Castrejon, Kaustav Kundu, Raquel Urtasun, and Sanja Fidler. Annotating object instances with a Polygon-RNN. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2017), pp. 5230–5238.

[176] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 616–625.

[177] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-BRS: Rethinking backpropagating refinement for interactive segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 8623–8632.

[178] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 5297–5306.

[179] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 577–585.

[180] Aida Syafiqah Ahmad Khaizi, Rasyiqah Annani Mohd Rosidi, Hong-Seng Gan, and Khairil Amir Sayuti. A mini review on the design of interactive tool for medical image segmentation. In: *International Conference on Engineering Technology and Technopreneurship* (2017), pp. 1–5.

[181] Risab Biswas. Polyp-sam++: Can a text guided sam perform better for polyp segmentation? In: *arXiv:2308.06623* (2023).

[182] Dhanush Babu Ramesh, Rishika Iytha Sridhar, Pulakesh Upadhyaya, and Rishikesan Kamaleswaran. Lung Grounded-SAM (LuGSAM): A Novel Framework for Integrating Text prompts to Segment Anything Model (SAM) for Segmentation Tasks of ICU Chest X-Rays. In: *Authorea Preprints* (2023).

[183] Jie Liu et al. Clip-driven universal model for organ segmentation and tumor detection. In: *IEEE/CVF International Conference on Computer Vision* (2023), pp. 21152–21164.

[184] Ziheng Zhao et al. One model to rule them all: Towards universal segmentation for medical images with text prompts. In: *arXiv:2312.17183* (2023).

[185] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. MedCLIP-SAM: Bridging text and image towards universal medical image segmentation. In: *arXiv:2403.20253* (2024).

[186] Marius George Linguraru et al. Tumor burden analysis on computed tomography by automated liver and tumor segmentation. In: *IEEE Transactions on Medical Imaging* 31.10 (2012), pp. 1965–1976.

[187] Myrthe AD Buser et al. Radiologic versus segmentation measurements to quantify Wilms tumor volume on MRI in pediatric patients. In: *Cancers* 15.7 (2023), p. 2115.

[188] Robert Seifert et al. Semiautomatically quantified tumor volume using 68Ga-PSMA-11 PET as a biomarker for survival in patients with advanced prostate cancer. In: *Journal of Nuclear Medicine* 61.12 (2020), pp. 1786–1792.

[189] Fei Lyu, Baoyao Yang, Andy J Ma, and Pong C Yuen. A segmentation-assisted model for universal lesion detection with partial labels. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2021), pp. 117–127.

[190] Philipp Seeböck et al. Exploiting epistemic uncertainty of anatomy segmentation for anomaly detection in retinal OCT. In: *IEEE Transactions on Medical Imaging* 39.1 (2019), pp. 87–98.

[191] Liang Jin et al. Interobserver agreement in automatic segmentation annotation of prostate magnetic resonance imaging. In: *Bioengineering* 10.12 (2023), p. 1340.

[192] Fiona R Kolbinger et al. Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise–an experimental study. In: *International Journal of Surgery* 109.10 (2023), pp. 2962–2974.

[193] Yannian Gu, Wenhui Lei, Hanyu Chen, Shaoting Zhang, and Xiaofan Zhang. Interactive Segmentation and Report Generation for CT Images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2025), pp. 273–283.

[194] Ruoqing Zhao, Xi Wang, Hongliang Dai, Pan Gao, and Piji Li. Medical report generation based on segment-enhanced contrastive representation learning. In: *CCF International Conference on Natural Language Processing and Chinese Computing* (2023), pp. 838–849.

[195] Mingyuan Meng et al. Adaptive segmentation-to-survival learning for survival prediction from multi-modality medical images. In: *NPJ Precision Oncology* 8.1 (2024), p. 232.

[196] Mingyuan Meng, Bingxin Gu, Lei Bi, Shaoli Song, David Dagan Feng, and Jinman Kim. DeepMTS: Deep multi-task learning for survival prediction in patients with advanced nasopharyngeal carcinoma using pretreatment PET/CT. In: *IEEE Journal of Biomedical and Health Informatics* 26.9 (2022), pp. 4497–4507.

[197] Vincent Andrearczyk et al. Multi-task deep segmentation and radiomics for automatic prognosis in head and neck cancer. In: *International Workshop on Predictive Intelligence in Medicine* (2021), pp. 147–156.

[198] K Harrison, H Pullen, C Welsh, O Oktay, J Alvarez-Valle, and R Jena. Machine learning for auto-segmentation in radiotherapy planning. In: *Clinical Oncology* 34.2 (2022), pp. 74–88.

[199] Tomaž Vrtovec, Domen Močnik, Primož Strojan, Franjo Pernuš, and Bulat Ibragimov. Auto-segmentation of organs at risk for head and neck radiotherapy planning: from atlas-based to deep learning methods. In: *Medical Physics* 47.9 (2020), e929–e950.

[200] Jordan Wong et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. In: *Radiation Oncology* 16.1 (2021), p. 101.

[201] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. In: *Nature Communications* 15.1 (2024), p. 654.

[202] Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. In: *European Conference on Computer Vision* (2024), pp. 207–229.

[203] Yufan He et al. VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025), pp. 20863–20873.

[204] Veronika Cheplygina, Adria Perez-Rovira, Wieying Kuo, Harm AWM Tiddens, and Marleen De Bruijne. Early experiences with crowdsourcing airway annotations in chest CT. In: *International Workshop on Deep Learning in Medical Image Analysis* (2016), pp. 209–218.

[205] Hao Tang et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. In: *Nature Machine Intelligence* 1.10 (2019), pp. 480–491.

[206] Pierre-Yves Baudin et al. A Minimal Annotation Pipeline for Deep Learning Segmentation of Skeletal Muscles. In: *NMR in Biomedicine* 38.7 (2025), e70066.

[207] Peter J Schüffler, Dig Vijay Kumar Yarlagadda, Chad Vanderbilt, and Thomas J Fuchs. Overcoming an annotation hurdle: Digitizing pen annotations from whole slide images. In: *Journal of Pathology Informatics* 12.1 (2021), p. 9.

[208] Karin Lindman, Jerómino F Rose, Martin Lindvall, Claes Lundstrom, and Darren Treanor. Annotations, ontologies, and whole slide images–development of an annotated ontology-driven whole slide image library of normal and abnormal human tissue. In: *Journal of Pathology Informatics* 10.1 (2019), p. 22.

[209] Zdravko Marinov, Paul F. Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen. Deep Interactive Segmentation of Medical Images: A Systematic Review and Taxonomy. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12 (2024), pp. 10998–11018.

[210] Adewunmi Akingbola, Oluwatimilehin Adeleke, Ayotomiwa Idris, Olajumoke Adewole, and Abiodun Adegbesan. Artificial intelligence and the dehumanization of patient care. In: *Journal of Medicine, Surgery, and Public Health* 3 (2024), p. 100138.

[211] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: *IEEE/CVF International Conference on Computer Vision* 2 (2001), pp. 416–423.

[212] Pedro RAS Bassi et al. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 15184–15201.

[213] Junlong Cheng et al. Interactive medical image segmentation: A benchmark dataset and baseline. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2025), pp. 20841–20851.

[214] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 11700–11709.

[215] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. In: *arXiv:1805.04398* (2018).

[216] Qin Liu, Zhenlin Xu, Gedas Bertasius, and Marc Niethammer. Simpleclick: Interactive image segmentation with simple vision transformers. In: *IEEE/CVF International Conference on Computer Vision* (2023), pp. 22290–22300.

[217] JunHao Liew, Yunchao Wei, Wei Xiong, Sim-Heng Ong, and Jiashi Feng. Regional interactive image segmentation networks. In: *IEEE/CVF International Conference on Computer Vision* (2017), pp. 2746–2754.

[218] Qin Liu, Jaemin Cho, Mohit Bansal, and Marc Niethammer. Rethinking interactive image segmentation with low latency high quality and diverse prompts. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 3773–3782.

[219] Soumajit Majumder and Angela Yao. Content-aware multi-level guidance for interactive instance segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 11602–11611.

[220] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022), pp. 1300–1309.

[221] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In: *IEEE International Conference on Image Processing* (2022), pp. 3141–3145.

[222] Hoang Le, Long Mai, Brian Price, Scott Cohen, Hailin Jin, and Feng Liu. Interactive boundary prediction for object selection. In: *European Conference on Computer Vision* (2018), pp. 18–33.

[223] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 13339–13348.

[224] Soumajit Majumder, Ansh Khurana, Abhinav Rai, and Angela Yao. Multi-stage fusion for one-click segmentation. In: *DAGM German Conference on Pattern Recognition* (2020), pp. 174–187.

[225] Zongyuan Ding, Tao Wang, Quansen Sun, and Fuhua Chen. Rethinking click embedding for deep interactive image segmentation. In: *IEEE Transactions on Industrial Informatics* 19.1 (2022), pp. 261–273.

[226] Arnaud Benard and Michael Gygli. Interactive video object segmentation in the wild. In: *arXiv:1801.00269* (2017).

[227] Henghui Ding, Scott Cohen, Brian Price, and Xudong Jiang. Phraseclick: toward achieving flexible interactive segmentation by phrase and click. In: *European Conference on Computer Vision* (2020), pp. 417–435.

[228] Daniel Beckmann, Jacqueline Kockwelp, Joerg Gromoll, Friedemann Kiefer, and Benjamin Risse. Sam meets gaze: Passive eye tracking for prompt-based instance segmentation. In: *NeuRIPS 2023 Workshop on Gaze Meets ML* (2023).

[229] Bin Wang, Armstrong Aboah, Zheyuan Zhang, Hongyi Pan, and Ulas Bagci. Gazesam: Interactive image segmentation with eye gaze and segment anything model. In: *Gaze Meets Machine Learning Workshop* (2024), pp. 254–265.

[230] Yinghuan Shi, Shu Liao, Yaozong Gao, Daoqiang Zhang, Yang Gao, and Dinggang Shen. Prostate segmentation in CT images via spatial-constrained transductive lasso. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2013), pp. 2227–2234.

[231] Yuxin Du, Fan Bai, Tiejun Huang, and Bo Zhao. Segvol: Universal and interactive volumetric medical image segmentation. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 110746–110783.

[232] Vien Ngoc Dang et al. Vessel-CAPTCHA: An efficient learning framework for vessel annotation and segmentation. In: *Medical Image Analysis* 75 (2022), p. 102263.

[233] Lena Heinemann et al. LIMIS: Towards Language-Based Interactive Medical Image Segmentation. In: *IEEE International Symposium on Biomedical Imaging* (2025), pp. 1–5.

[234] Feng Yang et al. Assessing inter-annotator agreement for medical image segmentation. In: *IEEE Access* 11 (2023), pp. 21300–21312.

[235] Stefan Jaeger et al. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. In: *Quantitative imaging in medicine and surgery* 4.6 (2014), p. 475.

[236] Feng Yang et al. Annotations of lung abnormalities in the Shenzhen chest X-ray dataset for computer-aided screening of pulmonary diseases. In: *Data* 7.7 (2022), p. 95.

[237] Xin Ye et al. Oimhs: An optical coherence tomography image dataset based on macular hole manual segmentation. In: *Scientific Data* 10.1 (2023), p. 769.

[238] Mario Amrehn et al. A Semi-Automated Usability Evaluation Framework for Interactive Image Segmentation Systems. In: *International journal of biomedical imaging* 2019.1 (2019), p. 1464592.

[239] Matthias Hadlich, Zdravko Marinov, Moon Kim, Enrico Nasca, Jens Kleesiek, and Rainer Stiefelhagen. Sliding window FastEdit: a framework for lesion annotation in whole-body pet images. In: *IEEE International Symposium on Biomedical Imaging* (2024), pp. 1–5.

[240] Aida Syafiqah Ahmad Khaizi, Rasyiqah Annani Mohd Rosidi, Hong-Seng Gan, and Khairil Amir Sayuti. A mini review on the design of interactive tool for medical image segmentation. In: *International Conference on Engineering Technology and Technopreneurship* (2017), pp. 1–5.

[241] Andriy Fedorov et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. In: *Magnetic Resonance Imaging* 30.9 (2012), pp. 1323–1341.

[242] Jun Ma et al. Efficient medsams: Segment anything in medical images on laptop. In: *arXiv:2412.16085* (2024).

[243] Geert Litjens et al. Evaluation of prostate segmentation algorithms for MRI: the PROMISE12 challenge. In: *Medical Image Analysis* 18.2 (2014), pp. 359–373.

[244] Federico Bolelli et al. Segmenting the inferior alveolar canal in cbcts volumes: the toothfairy challenge. In: *IEEE Transactions on Medical Imaging* 44.4 (2025), pp. 1890–1906.

[245] Daniele Falcetta et al. VesselVerse: A Dataset and Collaborative Framework for Vessel Annotation. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2025), pp. 655–665.

[246] Lena Maier-Hein et al. Metrics reloaded: recommendations for image analysis validation. In: *Nature Methods* 21.2 (2024), pp. 195–212.

[247] Oleksandr Kovalyk et al. PAPILA: Dataset with fundus images and clinical data of both eyes of the same patient for glaucoma assessment. In: *Scientific Data* 9.1 (2022), p. 291.

[248] Kai Jin et al. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. In: *Scientific Data* 9.1 (2022), p. 475.

[249] Yunpeng Wang et al. DeepSDM: Boundary-aware pneumothorax segmentation in chest X-ray images. In: *Neurocomputing* 454 (2021), pp. 201–211.

[250] Amir Hossein Abdi, Shohreh Kasaei, and Mojdeh Mehdizadeh. Automatic segmentation of mandible in panoramic x-ray. In: *Journal of Medical Imaging* 2.4 (2015), pp. 044003–044003.

[251] Walid Al-Dhabyani, Mohammed Gomaa, Hussien Khaled, and Aly Fahmy. Dataset of breast ultrasound images. In: *Data in Brief* 28 (2020), p. 104863.

[252] Debesh Jha et al. Kvasir-seg: A segmented polyp dataset. In: *International Conference on Multimedia Modeling* (2019), pp. 451–462.

[253] Hanna Piotrzkowska-Wróblewska, Katarzyna Dobruch-Sobczak, Michał Byra, and Andrzej Nowicki. Open access database of raw ultrasonic signals acquired from malignant and benign breast lesions. In: *Medical Physics* 44.11 (2017), pp. 6105–6109.

[254] Hanna Borgli et al. HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. In: *Scientific Data* 7.1 (2020), p. 283.

[255] Yunwei Hu et al. AMD-SD: An optical coherence tomography image dataset for wet AMD lesions segmentation. In: *Scientific Data* 11.1 (2024), p. 1014.

[256] Eva Blondeel et al. The Spheroid Light Microscopy Image Atlas for morphometrical analysis of three-dimensional cell cultures. In: *Scientific Data* 12.1 (2025), p. 283.

[257] Xin Zheng, Yong Wang, Guoyou Wang, and Jianguo Liu. Fast and robust segmentation of white blood cell images by self-supervised learning. In: *Micron* 107 (2018), pp. 55–71.

[258] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. In: *Scientific Data* 5.1 (2018), pp. 1–9.

[259] Noel Codella et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). In: *arXiv:1902.03368* (2019).

[260] Karen-Helene Støverud et al. AeroPath: An airway segmentation benchmark dataset with challenging pathology and baseline method. In: *Plos One* 19.10 (2024), e0311416.

[261] Rashed Karim et al. Evaluation of current algorithms for segmentation of scar tissue from late gadolinium enhancement cardiovascular magnetic resonance of the left atrium: an open-access grand challenge. In: *Journal of Cardiovascular Magnetic Resonance* 15.1 (2013), p. 105.

[262] Jin Ye et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. In: *arXiv:2311.11969* (2023).

[263] Diksha Goyal. Medical Image Segmentation Using Interactive Refinement. In: *Arizona State University* (2021).

[264] Theodora Kontogianni, Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Continuous adaptation for interactive object segmentation by learning from corrections. In: *European Conference on Computer Vision* (2020), pp. 579–596.

[265] Kun Li, George Vosselman, and Michael Ying Yang. Interactive image segmentation with cross-modality vision transformers. In: *IEEE/CVF International Conference on Computer Vision* (2023), pp. 762–772.

[266] Pushmeet Kohli, Hannes Nickisch, Carsten Rother, and Christoph Rhemann. User-centric learning and evaluation of interactive segmentation systems. In: *International Journal of Computer Vision* 100.3 (2012), pp. 261–274.

[267] Daniel Maleike, Marco Nolden, H-P Meinzer, and Ivo Wolf. Interactive segmentation framework of the medical imaging interaction toolkit. In: *Computer Methods and Programs in Biomedicine* 96.1 (2009), pp. 72–83.

[268] Junhyeok Lee, Han Jang, and Kyu Sung Choi. Domain-Specialized Interactive Segmentation Framework for Meningioma Radiotherapy Planning. In: *Workshop on Clinical Image-Based Procedures* (2025), pp. 32–41.

[269] Zhongzhen Huang, Yankai Jiang, Rongzhao Zhang, Shaoting Zhang, and Xiaofan Zhang. Cat: Coordinating anatomical-textual prompts for multi-organ and tumor segmentation. In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 3588–3610.

[270] Antonio Criminisi, Toby Sharp, and Andrew Blake. Geos: Geodesic image segmentation. In: *European Conference on Computer Vision* (2008), pp. 99–112.

[271] Pekka J Toivanen. New geodesic distance transforms for gray-scale images. In: *Pattern Recognition Letters* 17.5 (1996), pp. 437–450.

[272] Fredrik Andersson and Berit Kvernes. Bezier and B-spline Technology. In: *Umea University Sweden* (2003).

[273] Rolf Adams and Leanne Bischof. Seeded region growing. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 16.6 (1994), pp. 641–647.

[274] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas Huang. Deep grabcut for object selection. In: *arXiv:1707.00243* (2017).

[275] Camille Dupont, Yanis Ouakrim, and Quoc Cuong Pham. Ucp-net: Unstructured contour points for instance segmentation. In: *IEEE International Conference on Systems, Man, and Cybernetics* (2021), pp. 3373–3379.

[276] Daniel Kovacs-Deak, Maria Ines Meyer, Adriaan Lambrechts, Roel Wirix-Speetjens, and Frederik Maes. Design and Evaluation of Deep Learning Models for Interactive Image Segmentation: A Survey. In: *Authorea Preprints* (2025).

[277] Reza Azad et al. Medical image segmentation review: The success of u-net. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.12 (2024), pp. 10076–10095.

[278] Chun-Tse Lin, Wei-Chih Tu, Chih-Ting Liu, and Shao-Yi Chien. Interactive object segmentation with dynamic click transform. In: *IEEE International Conference on Image Processing* (2021), pp. 2284–2288.

[279] Sergios Gatidis et al. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. In: *Scientific Data* 9.1 (2022), p. 601.

[280] Abdelrahman Shaker, Muhammad Maaz, Hanoona Rasheed, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. UNETR++: delving into efficient and accurate 3D medical image segmentation. In: *IEEE Transactions on Medical Imaging* 43.9 (2024), pp. 3377–3390.

[281] Shehan Perera, Pouyan Navard, and Alper Yilmaz. Segformer3d: an efficient transformer for 3d medical image segmentation. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 4981–4988.

[282] Bambang Krismono Triwijoyo and Ahmat Adil. Analysis of medical image resizing using bicubic interpolation algorithm. In: *Jurnal Ilmu Komputer* 14.2 (2021), pp. 20–29.

[283] M Jorge Cardoso et al. Monai: An open-source framework for deep learning in healthcare. In: *arXiv:2211.02701* (2022).

[284] Ryosuke Okuta, Yuya Unno, Daisuke Nishino, Shohei Hido, and Crissman Loomis. CuPy: A NumPy-Compatible Library for NVIDIA GPU Calculations. In: *Workshop on Machine Learning Systems (LearningSys) in the Conference on Neural Information Processing Systems* (2017).

[285] Zdravko Marinov, Moon Kim, Jens Kleesiek, and Rainer Stiefelhagen. Rethinking Annotator Simulation: Realistic Evaluation of Whole-Body PET Lesion Interactive Segmentation Methods. In: *arXiv:2404.01816* (2024). To appear in ADSMI: MICCAI Workshop on Advancing Data Solutions in Medical Imaging AI, Marrakesh, Morocco.

[286] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman. Geodesic star convexity for interactive image segmentation. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2010), pp. 3129–3136.

[287] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: *International Conference on Machine Learning* (2016), pp. 1050–1059.

[288] Mario Amrehn, Maddalena Strumia, Markus Kowarschik, and Andreas Maier. Interactive neural network robot user investigation for medical image segmentation. In: *Bildverarbeitung für die Medizin 2019: Algorithmen–Systeme–Anwendungen. Proceedings des Workshops vom 17. bis 19. März 2019 in Lübeck* (2019), pp. 56–61.

[289] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, Sim-Heng Ong, and Jiashi Feng. Multiseg: Semantically meaningful, scale-diverse segmentations from minimal user input. In: *IEEE/CVF International Conference on Computer Vision* (2019), pp. 662–670.

[290] Zhen Xu et al. Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. In: *Patterns* 3.7 (2022), p. 100543.

[291] Jakob Wasserthal et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. In: *Radiology: Artificial Intelligence* 5.5 (2023), e230024.

[292] Junde Wu et al. Gamma challenge: glaucoma grading from multi-modality images. In: *Medical Image Analysis* 90 (2023), p. 102938.

[293] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. In: *IEEE Transactions on Medical Imaging* 23.4 (2004), pp. 501–509.

[294] Gongning Luo et al. Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound: The tdsc-abus challenge. In: *arXiv:2501.15588* (2025).

[295] Martina Melinščak, M Radmilovič, Zoran Vatavuk, and Sven Lončarić. Aroi: Annotated retinal oct images database. In: *International Convention on Information, Communication and Electronic Technology* (2021), pp. 371–376.

[296] Prasanna Porwal et al. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. In: *Data* 3.3 (2018), p. 25.

[297] Mingchao Li et al. OCTA-500: a retinal dataset for optical coherence tomography angiography study. In: *Medical Image Analysis* 93 (2024), p. 103092.

[298] Nikhila Ravi et al. Sam 2: Segment anything in images and videos. In: *arXiv:2408.00714* (2024).

[299] Shilong Liu et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: *European Conference on Computer Vision* (2024), pp. 38–55.

[300] Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for Medical Image Analysis. In: *Medical Image Analysis* 91 (2024), p. 102996.

[301] Josh Achiam et al. Gpt-4 technical report. In: *arXiv:2303.08774* (2023).

[302] Hugo Touvron et al. Llama: Open and efficient foundation language models. In: *arXiv:2302.13971* (2023).

[303] Aakanksha Chowdhery et al. Palm: Scaling language modeling with pathways. In: *Journal of Machine Learning Research* 24.240 (2023), pp. 1–113.

[304] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. In: *Nature Medicine* 29.8 (2023), pp. 1930–1940.

[305] Jianning Li, Amin Dada, Behrus Puladi, Jens Kleesiek, and Jan Egger. ChatGPT in healthcare: a taxonomy and systematic review. In: *Computer Methods and Programs in Biomedicine* 245 (2024), p. 108013.

[306] Jens Kleesiek, Yonghui Wu, Gregor Stiglic, Jan Egger, and Jiang Bian. An opinion on ChatGPT in health care—written by humans only. In: *Journal of Nuclear Medicine* 64.5 (2023), pp. 701–703.

[307] Rimsa Goperma, Rojan Basnet, Pragati Gautam Adhikari, Sagun Narayan Joshi, and Liang Zhao. NETRA: Enhancing Glaucoma Diagnosis Through Deep Learning-A Comparative Clinical Validation Study. In: *IEEE Region 10 Humanitarian Technology Conference* (2023), pp. 691–698.

[308] Tobias Friedetzki, Lorenz Haberzettl, Ricarda Buttmann, Frank Puppe, and Adrian Krenzer. iMedSTAM: Interactive Segmentation and Tracking Anything in 3D Medical Images and Videos. In: *CVPR 2025: Foundation Models for 3D Biomedical Image Segmentation* (2025).

[309] Yunyang Xiong et al. Efficient track anything. In: *IEEE/CVF International Conference on Computer Vision* (2025), pp. 11513–11524.

[310] Junwei Huang et al. autoPET IV challenge: Incorporating organ supervision and human guidance for lesion segmentation in PET/CT. In: *arXiv:2509.02402* (2025).

[311] Zhi Qin Tan, Xiatian Zhu, Owen Addison, and Yunpeng Li. U-mamba2: Scaling state space models for dental anatomy segmentation in cbct. In: *arXiv:2509.12069* (2025).

[312] Siddhartha Mallick, Jayanta Paul, and Jaya Sil. Response fusion attention U-ConvNext for accurate segmentation of optic disc and optic cup. In: *Neurocomputing* 559 (2023), p. 126798.