Deep Interactive Segmentation of Medical Images: A Systematic Review and Taxonomy

Zdravko Marinov, Paul F. Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen, *Member, IEEE*

***Abstract*—Interactive segmentation is a crucial research area in medical image analysis aiming to boost the efficiency of costly annotations by incorporating human feedback. This feedback takes the form of clicks, scribbles, or masks and allows for iterative refinement of the model output so as to efficiently guide the system towards the desired behavior. In recent years, deep learning-based approaches have propelled results to a new level causing a rapid growth in the field with 121 methods proposed in the medical imaging domain alone. In this review, we provide a structured overview of this emerging field featuring a comprehensive taxonomy, a systematic review of existing methods, and an in-depth analysis of current practices. Based on these contributions, we discuss the challenges and opportunities in the field. For instance, we find that there is a severe lack of comparison across methods which needs to be tackled by standardized baselines and benchmarks.**

***Index Terms*—Deep learning, interactive segmentation, medical imaging, systematic review.**

I. Introduction

Deep learning segmentation methods revolutionized various application areas including autonomous driving [158], product manufacturing [159], and medical image analysis [160]. For the latter, high-quality segmentation of anatomical structures and detection of abnormalities is an essential step towards automating diagnosis and treatment planning [161]. However, the quality of these methods relies heavily on large-scale data sets for training featuring high-quality annotations. Especially in the medical imaging domain, this poses a major bottleneck, because annotations are time-consuming and require expert knowledge [5]. For instance, labeling a volumetric Positron Emission Tomography/Computed Tomography (PET/CT) volume to identify tumor lesions can consume up to an hour of manual annotation for a single sample [154].

Deep interactive segmentation addresses this trade-off between high-quality segmentation and laborious manual annotation. The idea is to boost annotation efficiency by incorporating human feedback into either the training or application process of segmentation methods. This feedback loop lets users iteratively correct or refine the model output, e.g., in the form of clicks, scribbles, or fine-grained voxel-masks, and thus efficiently guide the model towards the desired output.

The development of interactive segmentation models in the medical domain entails unique challenges. Medical data is inherently diverse, including: 1) 2D images from dermoscopy, endoscopy, and microscopy; 2) 3D volumes from CT scans and other radiological sources; 3) and even videos from ultrasound. Thus, designing interactive segmentation models for specific modalities requires expertise in the best practices for that particular type of imaging. Moreover, achieving robust and accurate segmentation is further complicated by the variability introduced by different scanner types, population demographics, and the presence of noise and artifacts in the data [155].

The field of interactive segmentation traces back to active contour models [153] and Graph Cut [131], which primarily rely on low-level image features, such as pixel intensity changes, to differentiate foreground and background. However, traditional methods often require handcrafted features to incorporate high-level semantics related to the object-of-interest [5], [65], [140]. Additionally, traditional methods often require manual parameter tuning which can be domain or even image-specific [131], [141], [153]. These challenges have been widely solved in recent years by interactive deep learning-based approaches, as first introduced by Xu et al. [140].

Deep learning-based methods offer distinct advantages over traditional approaches by capturing both low- and high-level semantic features and being trained end-to-end without requiring image-specific parameter tuning. This paradigm shift has led to the successful application of interactive segmentation systems, for instance by reducing the annotation time of the aforementioned PET/CT volume to around three minutes [82].

Several reviews have been published in the field of interactive segmentation. However, previous reviews either focus on classical approaches rather than the more recent deep learning methods [145], [155], [156], [180], or exclude approaches from the medical domain [157]. At the same time, no review exists for the field of deep learning-based interactive segmentation of medical images despite its rapid emergence with over 121 proposed methods in the last 8 years as seen in Fig. 1. The lack of a systematic overview in this field hampers scientific

Received 21 November 2023; revised 15 July 2024; accepted 18 August 2024. Date of publication 30 August 2024; date of current version 5 November 2024. Recommended for acceptance by S.K. Zhou. *(Zdravko Marinov and Paul F. Jäger are co-first authors.) (Corresponding author: Zdravko Marinov.)*

Zdravko Marinov and Rainer Stiefelhagen are with the Computer Vision for Human-Computer Interaction Lab, Department of Informatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany (e-mail: zdravko.marinov@kit.edu).

Paul F. Jäger is with the German Cancer Research Center (DKFZ) Heidelberg, Interactive Machine Learning Group, 69120 Heidelberg, Germany, and also with Helmholtz Imaging, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany.

Jan Egger and Jens Kleesiek are with the Institute for Artificial Intelligence in Medicine (IKIM), University Hospital Essen (AöR), 45131 Essen, Germany.

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2024.3452629, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2024.3452629

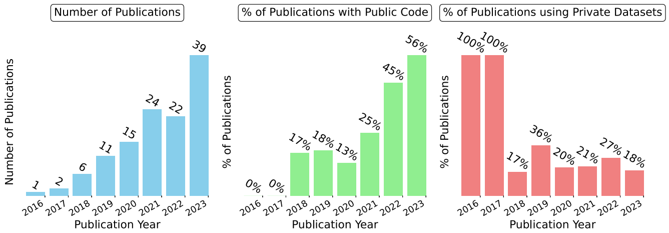


Fig. 1. Tendencies in medical interactive segmentation in recent years.

progress by generating redundancies and poses a challenge for users seeking the best-fitting method for their problem.

We address these shortcomings in this dedicated review by means of the following key contributions:

- We introduce a comprehensive taxonomy for deep interactive segmentation allowing users to quickly comprehend the various approaches and select the best fitting method for their task.
- Based on this taxonomy, we provide a systematic review of 121 proposed methods in the medical domain.
- We perform an in-depth analysis of the current practices in the field including prevalent datasets, anatomies, and validation metrics, as well as the adequacy of baselines and the reproducibility of results.
- Based on this analysis, we provide a discussion of current challenges and opportunities in the field.

II. Terminology

Before we present our systematic review, we establish clear definitions for the fundamental terminology within the domain of interactive segmentation.

A. *Interactive Segmentation*

Interactive segmentation describes an iterative feedback loop, where user-provided corrections or refinements to the model's output inform subsequent iterations, leading to updated predictions. Depending on the method, user guidance is provided during training or application in the form of, e.g., clicks, scribbles, or other interactions. Importantly, initial labels provided to a model before training are excluded from this definition to differentiate interactive segmentation from related training paradigms such as weakly-supervised segmentation.

B. *Guidance Signal*

A guidance signal is a representation of the user interactions in a form in which the model can process it. This can be an explicit representation that involves transforming the user interaction into an additional structured input for the model to process and learn from, e.g., Gaussian heatmaps centered around user clicks. Additionally, guidance signals can also be implicit, where user interaction information is subtly integrated into the model's learning process without the provision of explicit structured input. For instance, this integration could involve modifying the loss function to incorporate the distance to clicks and assign greater weight to predictions in proximity to those clicks. Existing guidance signals for clicks, scribbles, and other interactions are given in the Appendixes.

C. *Training and Application*

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.

D. *Robot User*

The concept of a robot user [122] involves creating a simulated model that mimics the behavior of a real human annotator. The robot user leverages ground-truth labels to simulate user interactions at plausible locations. For example, clicks can be sampled randomly from the ground-truth labels or generated at the center of the largest object. These simulated interactions are then converted into a guidance signal, which is fed back to the model. Robot users are used during training to simulate interactions for a large number of training samples as this is unfeasible for real human annotators at this scale. Additionally, robot users can also be used during application to evaluate trained models on unseen data without involving real human annotators. Robot users can be categorized as non-iterative or iterative. Non-iterative 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 [152]. Here, an iteration denotes a single round of interaction and prediction with the model.

III. Scope and Study Collection Strategy

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 the PRISMA 2020 checklist in the Appendixes. We performed a literature search in several databases, including PubMed, Google Scholar, IEEE Xplore, SpringerLink, and arXiv, 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.

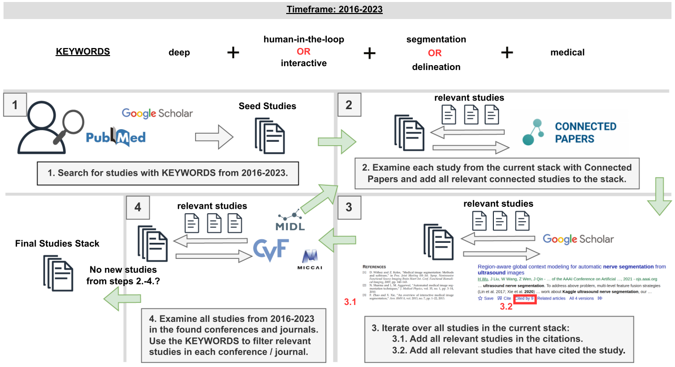

Timeframe: 2016-2023
KEYWORDS
deep
+
human-in-the-loop
OR
interactive
+
segmentation
OR
delineation
+
medical
1
Google Scholar
PubMed
Seed Studies
1. Search for studies with KEYWORDS from 2016-2023.
2
relevant studies
CONNECTED PAPERS
2. Examine each study from the current stack with Connected Papers and add all relevant connected studies to the stack.
3
relevant studies
Google Scholar
References
Region-aware global context modeling for automatic nerve segmentation from ultrasound images
Save
Cite
Cited by 9
Related articles
All 4 versions
3.1
3.2
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
relevant studies
MIDL
CVF
MICCAI
4. Examine all studies from 2016-2023 in the found conferences and journals. Use the KEYWORDS to filter relevant studies in each conference / journal.
Final Studies Stack
No new studies from steps 2.-4.?

Fig. 2. 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 is given in the Appendixes.

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 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.

One reviewer assessed a study's eligibility through a three-stage process. Initially, we examine the title to decide if it focuses on deep medical interactive segmentation. If the title is ambiguous, we read the abstract for confirmation. In cases where the abstract remains unclear, we read the entire study.

This search produced our initial *seed studies* stack as illustrated in Fig. 2. 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 4in Fig. 2. In step 2, we incorporated the Connected Papers tool[1] 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 meet our pre-defined 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, which is given in the Appendixes, and manually screened all of the publications from each venue in the timeframe 2016–2023 with our pre-defined 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, one reviewer 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 VIII and IX in the Appendixes. This facilitates efficient navigation for future researchers seeking relevant interactive methods related to their own work.

It is important to note that during our search we exclude "classical approaches", which do not utilize deep learning. Some examples include methods based on: 1) Graph Cut [131], [141]; 2) dense Conditional Random Fields (CRFs) [166]; 3) active contours [153]; and 4) level sets [167]. While non-deep learning interactive frameworks such as ilastik [168] and ITK-Snap [169] have demonstrated success in clinical workflows, we maintain a focus on deep learning-based methods to align with the review's scope and the growing prevalence of interactive deep learning models in the medical domain.

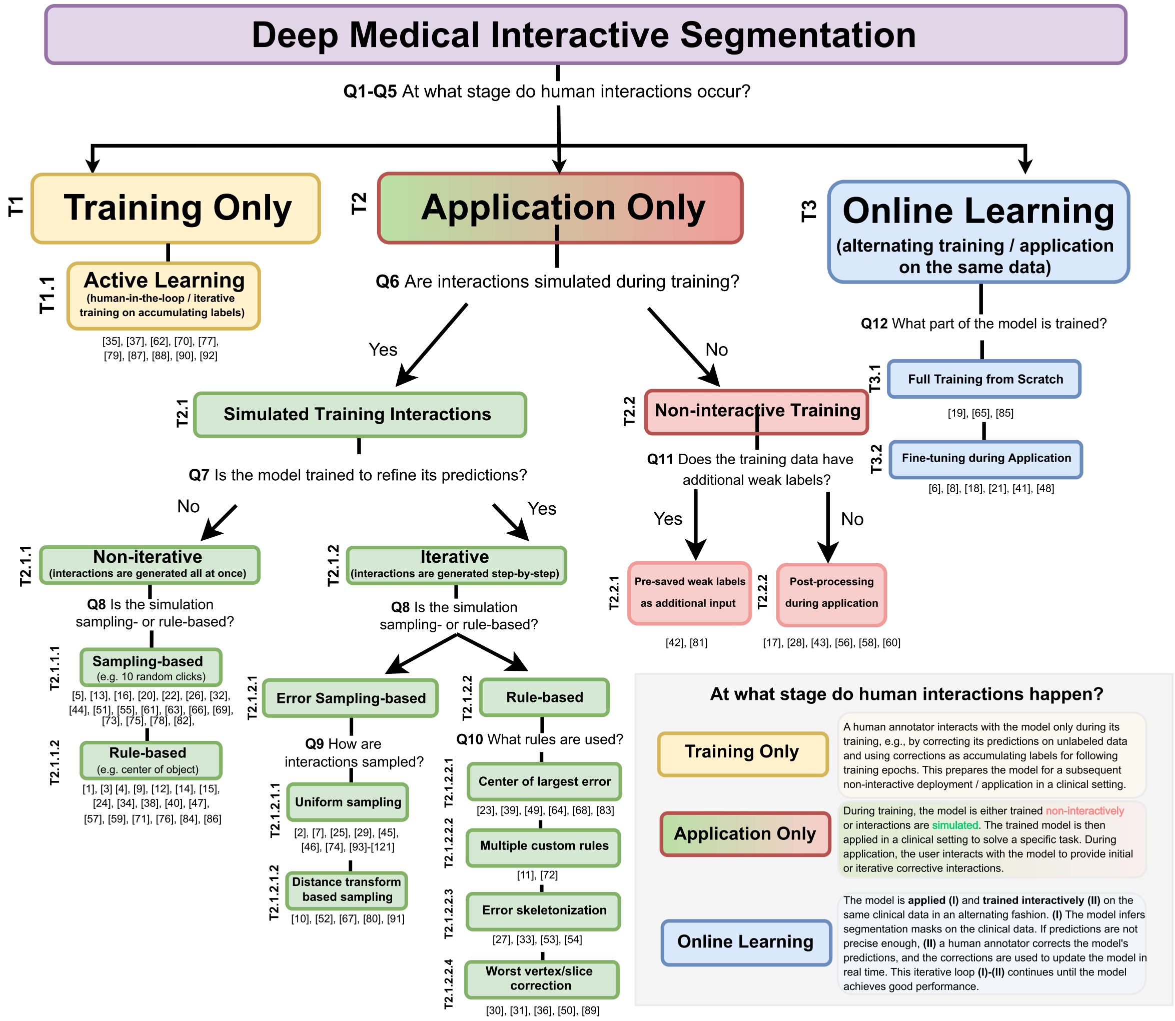
IV. Taxonomy

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* and *taxonomy blueprints* in Figs. 3 and 5, which function as navigational tools for existing medical interactive segmentation methods. These tools 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 both tools.

Taxonomy tree: In our systematic review of deep medical interactive segmentation, 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 Fig. 3, and a summary of each paradigm can be found in the three boxes at the bottom right. The interactions take place in two distinct stages: training and application, which are defined in Section II-C. 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 IV-A, IV-B, and IV-C.

Taxonomy blueprints: Fig. 5 visually depicts 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. The distinction between training and application phases is apparent in most paradigms, however, in the case of the *online learning* paradigm, this separation is not as evident. In *online learning*, the model is alternately trained and applied to the same data, with real-time feedback provided

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

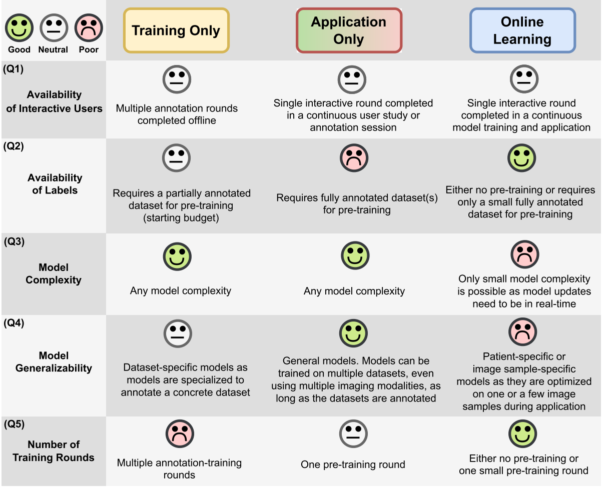


A human annotator interacts with the model only during its training, e.g., by correcting its predictions on unlabeled data and using corrections as accumulating labels for following training epochs. This prepares the model for a subsequent non-interactive deployment / application in a clinical setting.

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.

Fig. 3. Our proposed taxonomy tree for all the reviewed studies. The references for studies associated with a node are listed beneath the respective node.



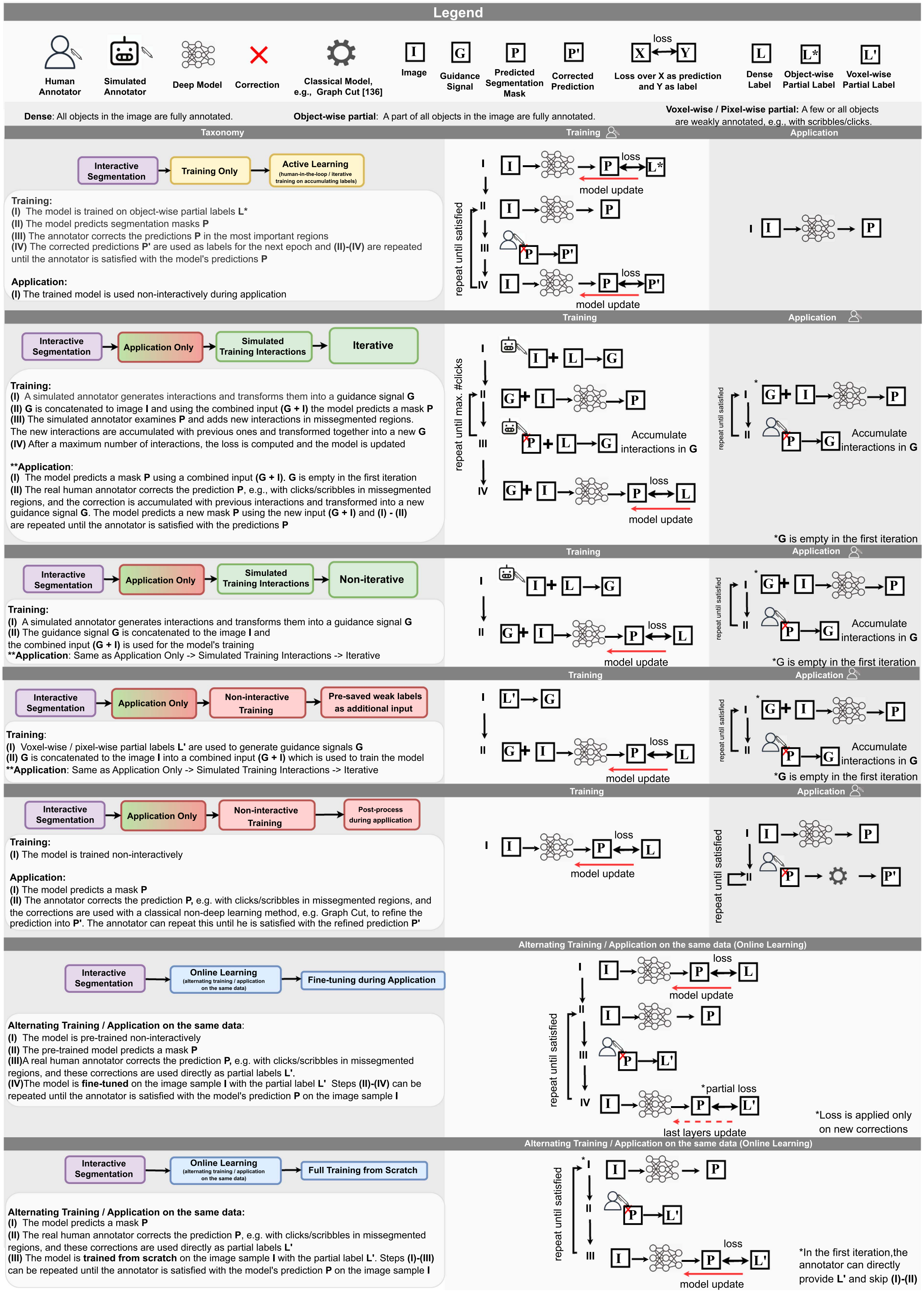
🙂 Good 😐 Neutral 🙁 Poor	Training Only	Application Only	Online Learning
(Q1) Availability of Interactive Users	😐 Multiple annotation rounds completed offline	😐 Single interactive round completed in a continuous user study or annotation session	😐 Single interactive round completed in a continuous model 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

Fig. 4. Advantages and disadvantages of methods within the three main taxonomy nodes regarding five specific demands.

by a human annotator. The taxonomy blueprints offer a two-fold advantage: 1) they reveal detailed differences in training and application phases among nodes in the taxonomy tree; 2) and streamline the categorization of emerging methods. They serve as a visual guide to both understand the taxonomy nodes and systematically incorporate new approaches into the existing taxonomy structure.

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 Fig. 4 and present the questions in the following.



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.

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

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**

*G is empty in the first iteration

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

*G is empty in the first iteration

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

*G is empty in the first iteration

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'**

Alternating Training / Application on the same data (Online Learning)
Interactive Segmentation
Online Learning
(alternating training / application on the same data)
Fine-tuning during Application
repeat until satisfied
I
II
III
IV
loss
model update
*partial loss
last layers update

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**

*Loss is applied only on new corrections

Alternating Training / Application on the same data (Online Learning)
Interactive Segmentation
Online Learning
(alternating training / application on the same data)
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)**

Fig. 5. Taxonomy blueprints for our proposed taxonomy nodes. The human annotator is involved during training, application, or both in *online learning*.

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 Fig. 4 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 Fig. 3) or non-interactive pre-training (red branch in Fig. 3). *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.

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.

A. 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 Fig. 3. 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 Fig. 5. Ho et al. [35] introduce active learning to deep medical interactive segmentation, utilizing a Convolutional Neural Network (CNN) on an unlabeled osteosarcoma dataset, substantially reducing annotation time compared to pixelwise annotation. Menon et al. [37] introduce a method where annotators highlight a query patch for annotating whole-slide images (WSIs). A retrieval module selects the K-nearest patches based on the feature space similarity and annotators offer feedback for each patch as relevant or irrelevant or provide explicit segmentation labels. Using their retrieval module, only 5% of patches need annotation for state-of-the-art performance. Atzeni et al. [62] present a method leveraging estimated segmentation quality and labeling effort to identify regions of interest. The labeling effort considers boundary length and irregularity, assuming complex boundaries are harder to annotate. The segmentation quality is measured by the average class Dice score on annotated regions. This selects easier regions for initial annotation rounds, aiding the model in learning valuable features for annotating more challenging areas later. AnatomySketch [70] presents an open-source software platform with a graphical user interface designed for annotating and integrating deep learning segmentation models. The "Annotation-by-iterative-Deep-Learning (AIDL)" module enables annotators to proofread, correct, and incorporate segmentations into the next training iteration of a pre-trained model. Deep SED-Net [77] shows that an AIDL strategy for testicular cell segmentation achieves results comparable to manual annotation using squeeze-and-excitation layers [130] in a U-Net model [127]. Ma et al. [79] use an igniter network trained on a small dataset to generate coarse labels for a larger dataset, which are then refined by a human annotator in an AIDL loop following a specific labeling protocol. This protocol prioritizes easier samples for early labeling and gradually addresses harder ones, minimizing human effort while enhancing the model's predictions, similar to Atzeni et al. [62]. Zhuang et al. [87] propose a boundary contour correction tool as an alternative to voxel-wise corrections, showing enhanced shape learning, faster proofreading, and more anatomically plausible results. Ho et al. [88] expedite the AIDL paradigm by utilizing a pre-trained breast segmentation model instead of random weight initialization, decreasing the annotation time. Zhuang et al. [90] employ user-provided scribbles to compute an exponentialized geodesic distance map, used to modulate the model's prediction and generate pseudo-labels for the next training iteration. These pseudo-labels are more certain near the scribbles and integrate human feedback during training. Qu et al. [92] train U-Net [127], Swin-UNETR [135], and nnU-Net [136] on a small CT dataset, then use their predictions to annotate 8000 CT volumes. They assess prediction inconsistency, entropy, and overlap to suggest volumes for refinement to annotators, reducing the annotation time to two work weeks.

B. 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 Fig. 3. 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:

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.

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 Fig. 5. 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 "behavior gap" 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.

T2.1.1.1. Non-iterative sampling-based methods sample the ground-truth labels to simulate interactions. DeepIGeoS [5] samples a fixed amount of voxels from connected components that are over a certain size threshold and uses them as seeds for computing a geodesic distance transform. UGIR [22], Bi et al. [63], DeepIGeoSv2 [13], WDTISeg [44], and Hallitschke et al. [82] follow the same sampling strategy as DeepIGeoS [5]. UGIR [22] additionally estimates the segmentation uncertainty by calculating the prediction variance within a group convolution layer. Bi et al. [63] integrate the guidance signal at multiple stages in their skin lesion segmentation model. DeepIGeoSv2 [13] expands upon the two-stage DeepIGeoS [5] model to handle multiple organs and introduces an uncertainty-aware loss function that assigns an exponential penalty based on the model's certainty of an error. WDTISeg [44] combines geodesic and euclidean distance maps through a linear combination, allowing the incorporation of both appearance and location information, respectively. Hallitschke et al. [82] expand DeepIGeoS [5] to multimodal PET/CT data, exploring various annotation interface presentations for users when displaying multimodal data. Cerrone et al. [16] segment neuron cells from serial section electron microscopy images by randomly sampling a click from each neuron while maintaining a minimum distance from any boundary. Wang et al. [20] perturb ground-truth polygon vertices by applying randomly sampled offsets and directions. NuClick [26] randomly samples a point within each nucleus, ensuring it is at least two pixels away from the object boundaries. Tang et al. [32] dilate ground-truth masks of liver and lung lesions, along with lymph nodes, then randomly sample five pixels from the dilated mask. Jiang et al. [51] use a two-stage network and randomly sample clicks from the segmentation error of the first-stage coarse network, encoding them as Gaussian heatmaps. Daulatabad et al. [55] sample multiple clicks in the proximity of the centroid of the ground-truth mask of the thyroid nodule. Shi et al. [69] partition the ground-truth mask based on the distance to the object's boundary, then randomly sample one pixel from each section, addressing cluttered samples in the guidance map. Shahedi et al. [61] and Ju et al. [78] uniformly sample clicks from the target organ, varying the number of sampled clicks in their ablation studies. Pirabaharan et al. [73], [75] uniformly sample the ground-truth mask to generate foreground and background clicks. These clicks are encoded as Gaussian heatmaps, with a radius proportional to the mask's area. Smaller radii for smaller objects ensure better boundary alignment.

T2.1.1.2. Non-iterative rule-based methods employ deterministic rules to simulate interactions. Sun et al. [3] simulate a click at the prostate center. They use Canny edge detection [123] to create horizontal and vertical location prior maps, assigning decreasing intensity to voxels farther from the central click with more crossed edges. Khan et al. [12] use object extreme points (topmost, leftmost, rightmost, bottom-most) as four clicks and generate a confidence map based on Chebyschev and Mahalanobis distances to the object center. DeepCut [1] extends the ground-truth bounding box to generate foreground and background voxels for an interactive CNN. Dense CRFs [166] refine the CNN's predictions, and then the refined predictions are used as foreground and background voxel seeds for the interactive CNN once more. Can et al. [4] also utilize dense CRFs to refine CNN predictions for prostate and cardiac structures segmentation. iW-Net [14] simulates two clicks by using the farthest points in the ground truth mask to compute an attraction field, inspired by oppositely charged punctual electric charges. Roth et al. [15], [40] use 3D Gaussian heatmaps centered at extreme points, expanded and employed as guidance for a CNN-based model. Raju et al. [24] bridge the domain gap between simulated and ground-truth extreme points by training a model to predict them on unseen data, then use them as a guidance signal. Girum et al. [34] use extreme points as input to their prior-knowledge network. This network produces a spatial attention map, which is multiplied with the image and fed into a downstream segmentation model. MIDeepSeg [38] uses extreme points to simulate clicks, slightly shifting them inward to obtain interior margin points. These points are then used to compute an exponentialized geodesic distance map as a guidance signal. Zhang et al. [47] extract image patches along rays from the object's center to its outer boundaries to train a Convolutional Recurrent Neural Network (ConvRNN). During application, a single click at the object's center is enough, as the ConvRNN segments neighboring patches around the click sequentially. Trimpl et al. [57] simulate full annotation on the central axial slice of a CT image and propagate it to the remaining slices. Their model utilizes the central slice, its annotation, and a target slice as joint inputs to learn the segmentation of the target slice. Iteratively selecting each slice as a new target slice segments the entire volume. Jahanifar et al. [59] skeletonize the ground-truth mask to simulate scribbles. i3Deep [66] randomly samples multiple slices from ground-truth labels per image sample. The full image labels of selected slices are appended to the input of a refinement model during training. Galisot et al. [71] train segmentation models on various brain structures using cropped brain regions as inputs. They also develop a model to learn spatial relationships between structures, automatically positioning bounding boxes during inference, with annotators able to adjust them as needed. Chen et al. [76] generate 2D Gaussian heatmaps around each extreme point and compute an euclidean distance transform using the intersection point of the two extreme axes as a seed point for a second guidance signal. Bruzadin et al. [84] propagate foreground seeds from a source slice to neighboring slices by considering strong edges in the image and avoiding sampling seeds near those edges in adjacent slices. Shahin et al. [86] identify the slice with the highest error and use its ground-truth boundary as a scribble.

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. 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.)***.

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. Uniform sampling methods are more computationally efficient than distance transform-based ones. However, distance transforms enable sampling closer to the central regions of the label.

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.

T2.1.2.1.1. Iterative uniform sampling approaches sample new interactions with an equal probability of landing in any of the missegmented pixels/voxels. UI-Net [2] detects missegmented regions of hepatic lesions and samples a random number of pixels for each interaction. For the first interaction, they initialize foreground and background scribbles by applying multiple dilation and erosion operations on the lesion's boundary. InterCNN [7] uniformly samples multiple clicks from the error, places a 9×9 window around each click, and adds all foreground pixels in the window to the sampled clicks, regardless of whether they were missegmented. Hu et al. [29] use a stratification approach, randomly sampling a click from each of the three largest missegmented connected components. Li et al. [45] randomly sample clicks from the intersection of the object's boundary and error regions. They use a reinforcement learning approach, rewarding the agent based on cross-entropy improvement. Additionally, they propose a confidence estimation network guiding annotators by suggesting click locations based on segmentation confidence. Deng et al. [46] use a sampling strategy selecting a fixed number of under- and oversegmented voxels per iteration. Their loss function targets only the $9 \times 9 \times 9$ neighborhood around each missegmented voxel to avoid affecting well-segmented regions. Mikhailov et al. [74] randomly sample clicks from missegmented regions, storing them in an ordered memory bank across all iterations. This preserves the sequence of interactions, ensuring the sequential information is retained instead of combining all clicks into a single guidance signal.

Recently, Meta AI released the code for their Segment Anything Model (SAM) [137]. Due to its remarkable performance and zero-shot capabilities on natural images, many methods have adapted SAM for medical images. In this review, we only consider methods that use SAM's interactive prompts. All the reviewed medical SAM methods fall into the category of iterative uniform sampling simulation in our taxonomy, which uses SAM's original pre-training described in the training algorithm in [137, p.17]. Here, we summarize these methods.

Mazurowski et al. [93] extensively evaluate SAM's zero-shot performance on 33 datasets, exhibiting significant performance variations across tasks, ranging from 0.11 to 0.86 Intersection over Union. They find bounding box prompts consistently yield superior results, and that SAM performs better on larger objects. Iterative corrections do not lead to substantial improvements, with the best performance achieved in the first three clicks for most tasks. Deng et al. [94] find SAM excels in segmenting larger objects but struggles with multiple small objects, even with abundant prompts. The study concludes SAM is unsuitable for gigapixel whole-slide imaging (WSI) data. SAM vs. BET [95] demonstrates SAM's superiority over the gold standard Brain Extraction Tool (BET) [134] in brain extraction from MRI images, however, it does not compare it to newer skull stripping models [101]. Putz et al. [96] show SAM's effective generalization in glioma brain tumor segmentation, except for small tumors under 300mm^3 , where performance deteriorates slightly. Hu et al. [97] assess SAM's effectiveness in liver tumor segmentation, finding a significant performance gap compared to even a simple U-Net model [127]. SAM-Adapter [98] enhances SAM by injecting task-specific embeddings into its image encoder, resulting in significant performance improvement for polyp segmentation compared to using SAM directly without modifications. Ophthalmology SAM [100] fine-tunes SAM with an additional prompt adapter on fundus images and improves SAM significantly on three ophthalmology tasks.

He et al. [101] assess SAM on 12 public medical datasets across ten organs and six imaging modalities. They find SAM is consistently outperformed by a simple U-Net [127] across all datasets, with performance strongly influenced by the target object's size. Additionally, SAM achieves higher results on 2D modalities (dermoscopy, colonoscopy, X-Ray, ultrasound) compared to 3D modalities like MRI and CT. Shi et al. [102] confirm SAM's inferiority to a basic U-Net model [127] in fundus, CT, MRI, and Optical Coherence Tomography (OCT) data. However, they demonstrate that in-domain fine-tuning enables SAM to achieve specialized U-Net model performance in retinal vessel segmentation. GazeSAM [103] employs eye gazing to estimate the annotator's point of focus, encodes this position as a click, and utilizes SAM for segmentation. Skin-SAM [104] fine-tunes SAM on dermoscopy images by using simulated bounding box prompts, resulting in satisfying performance on skin lesion segmentation. Wang et al. [105] employ SAM for surgical instrument segmentation, finding bounding box prompts significantly outperform click prompts. However, SAM's performance remains unsatisfactory in challenging scenarios, such as overlapping instruments and blood. Cheng et al. [106] evaluate SAM without fine-tuning on 12 medical datasets, showing bounding boxes yield notably better results than clicks. Additionally, they find incorporating perturbations into bounding boxes decreases performance. Mattjie et al. [107] explore SAM across six datasets, confirming that using the ground-truth bounding box without perturbations consistently yields optimal results across all datasets and transformer backbones. Polyp-SAM [108] fine-tunes SAM on five colonoscopy datasets with bounding box prompts. They discover that fine-tuning solely the decoder and using a smaller transformer backbone yields the best performance. BreastSAM [110] also concludes that using a smaller transformer backbone leads to better results for breast cancer segmentation in ultrasound images. IAMSAM [111] implements an annotation interface for microscopy images, using segmentation masks for downstream tasks like cell type prediction and spatial transcriptomics. Shen et al. [113] expand SAM with temporal prompts, where a Reinforcement Learning (RL) agent advises the suitable prompt type, like a bounding box or click. Their study shows RL suggestions outperform choosing a single interaction type. Ning et al. [114] utilize SAM on ultrasound videos, revealing its potential for segmenting diverse structures, with minimal deviations between frames when enough prompts are provided. Zhang et al. [115] experiment across various anatomical structures, finding SAM excels on large organs like the liver and brain. Yet, its performance declines for smaller and ambiguous targets like the parotid and cochlea. MedLSAM [116] uses extreme points that implicitly define a bounding box prompt for SAM and reduce the annotation burden. SAM-U [117] produces multiple bounding box prompts for a single image. It estimates the aleatoric uncertainty by computing the prediction entropy from separate forward passes with each bounding box. This metric identifies challenging regions requiring further annotator guidance. 3DSAM-adapter [118] adapts SAM to 3D images and prompts by freezing pre-trained weights and extending SAM's components to 3D. The patch embedding is extended with a 3D depth-wise convolution, and the 3D position encoding integrates the original 2D lookup table with a new depth lookup table. The attention block queries expand from $[B, H \times W, c]$ to $[B, H \times W \times D, c]$, and all 2D convolutions are replaced with 3D in the bottleneck. Huang and Yang et al. [119] evaluate SAM across 52 public datasets, exploring its "Segment Everything" mode and various click and bounding box prompts. They find SAM's performance varies significantly across datasets and modalities. Bounding boxes consistently outperform clicks, while the "Segment Everything" mode performs the worst.

In contrast to the predominantly negative findings in most other works that integrate SAM for medical images, MedSAM [120] achieves a remarkable performance on 14 unseen datasets, covering 50 target classes and seven imaging modalities, and even surpasses specialized nnU-Net [136] models on each of the target classes. This impressive outcome is the result of the careful curation of 84 existing public medical datasets for pre-training, leading to 1 090 486 medical image-mask pairs, and fine-tuning SAM on this large-scale medical dataset. The diversity of this dataset, spanning 15 imaging modalities, bolsters MedSAM's strong generalization abilities and reveals the significant potential of using SAM for medical interactive segmentation. Furthermore, MedSAM [120] concludes that bounding box prompts perform the best, and they convert 3D images into 2D slices for training and evaluation. Medical Sam Adapter (MSA) [99] extends depth attention to address the dimensionality reduction from 3D to 2D images in SAM's training. Pre-training on large-scale medical datasets, MSA exhibits superior performance to MedSAM [120], but only when using clicks instead of bounding boxes. PromptUNet [109] train an interactive one-prompt model on 64 medical datasets, surpassing both click-based MedSAM [120] and MSA [99] on 14 datasets. DeSAM [112] separates the prompt from the image to mitigate the impact of inadequate prompts, significantly enhancing performance over MedSAM [120], even with bounding box prompts.

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. Sakinis et al. [10] utilize the Chamfer distance transform on errors, employing the resulting distance map as a sampling distribution for new clicks. Bai et al. [52] utilize the euclidean distance transform on over- and under-segmented regions, converting them into background and foreground sampling distributions. They then exponentiate and normalize the distance maps to obtain pseudo-probability maps. DeepEdit [67] follows Sakinis et al. [10], while also exploring varied proportions of interaction-free iterations, where the model receives no clicks (an empty guidance signal). Bai et al. [80] use the euclidean distance transform on error regions, followed by Softmax normalization to generate a pseudo-probability map for sampling new clicks. Guiding the Guidance (GtG) [91] builds upon DeepEdit [67] by introducing a dynamic Gaussian heatmap with varying radii. They assess four guidance signals and introduce five metrics for a comprehensive evaluation of interactive models.

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: 1) center of largest error ***(T2.1.2.2.1.)***; 2) error skeletonization ***(T2.1.2.2.2.)***; 3) multiple custom rules ***(T2.1.2.2.3.)***; and 4) worst vertex/slice correction ***(T2.1.2.2.4.)*** and 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 (clicks, scribbles, etc.).

T2.1.2.2.1. Center of largest error methods use the center of the largest error region as the next click with the assumption that it is the most intuitive choice. IterMRL [23] and BS-IRIS [25] utilize multi-agent reinforcement learning, treating each voxel as an agent with a cross-entropy improvement reward. IterMRL [23] selects the center of the largest error region and $N - 1$ other largest connected components in each iteration. Feng et al. [39] merge few-shot learning with interactive segmentation by training with annotations on a small set of slices, clicking only on those slices. Subsequent clicks are positioned at the center of the largest connected error component within these slices. DINs [49] place clicks at the center of the largest error region, verifying if it matches the ground-truth class. In concave regions, they skeletonize the error region and select the nearest point in the skeleton as the click location, ensuring precise placement. iSegFormer [64] samples clicks at the centers of under- and oversegmented regions for knee cartilage segmentation. Liu et al. [68] employ a transformer-based model for multiclass segmentation, targeting missegmented regions by placing clicks at their centers for each class. Liu et al. [83] maintain initial segmentation quality via cycle consistency, starting with a click in the largest structure and refining the worst-segmented organ using central clicks iteratively.

T2.1.2.2.2. Error skeletonization 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. A visual example of error skeletonization is given in the Appendixes. Kitrungrotsakul et al. [27] employ skeletonization to generate foreground and background scribbles from under- and oversegmented errors. These scribbles are utilized by a second-stage model, augmenting the initial non-interactive model. Jinbo et al. [33] extend Kitrungrotsakul et al.'s [27] method with an improved annotation interface enabling simultaneous scribble drawing and real-time segmentation visualization. DeepScribble [53] generates scribble-like annotations by thresholding and skeletonizing euclidean distance maps computed for false positive and false negative regions. Attention-RefNet [54] leverages skeletonized errors to mimic scribbles, generating a guidance signal by subtracting geodesic distance maps of the foreground and background, assigning positive values to the foreground and negative values to the background.

T2.1.2.2.3. Multiple custom rules are methods that apply multiple custom rules that are specific to the application. Zhou et al. [11] simulate central clicks by eroding the largest connected component and selecting the center from the remaining pixels. They also simulate scribbles by connecting the two farthest points in the largest error component of the worst segmented 2D slice. Lin et al. [72] simulate a boundary scribble by dilating the object's boundary and simulate iterative clicks by placing a central click in the largest error region.

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. Tian et al. [30] predict boundary polygon vertices and mimic user interaction by adjusting the worst vertex toward its correct position. They employ a Graph Convolutional Network (GCN) to propagate adjustments to the remaining vertices and update the segmentation contour. Tian et al. [50] expand this method with a local correction, adjusting only the $2 \times K$ neighboring vertices of the corrected vertex. This local update preserves well-segmented regions while refining local errors. Foo et al. [36] find the slice with the largest error and find the pair of points that is the furthest apart and connect them to form the scribble. Chao et al. [31] target the most poorly segmented 2D slice, using its corrections to update the model's bottleneck features. This iterative refinement process enhances the segmentation model's performance across the entire volume. Wei et al. [89] also simulate a slice correction and use the ground-truth label of the slice with the largest tumor area as a guidance signal. They compare this strategy to random slice corrections to show that their approach selects more informative slices.

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 Fig. 3. 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.

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. Zhou et al. [42], [81] utilize weak labels in the form of scribbles on a single source slice and propagate the label information from the source slice to the rest of the volume using a memory-readout operation from a memory-encoder network.

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. For instance, Zheng et al. [17] use shadow set theory [124] to extract ground-truth masks from the training dataset. These masks are aligned with human clicks on unseen images during application and the averaged extracted masks are fused with the model's prediction, with the mask variance estimating the uncertainty. In IRIS [28], patches around user clicks are extracted and fed into a pre-trained model. Post-processing stitches together predictions from all patches to obtain the final prediction. Williams et al. [43] use B-spline active surfaces [125] to calculate contours along CNN predictions. Users modify contours by dragging control points, updated via Yezzi energy minimization [126]. PiPo-Net [58] employs a two-stage model: a U-Net [127] generates pixelwise masks, and an LSTM [128] produces a vertex polygon. Users correct vertices, updated by the LSTM as a post-processing step. Manh et al. [56] employ a U-Net [127] for Z-line segmentation, followed by Binary Partition Tree (BPT) [132] post-processing. Users mark superpixels [133] with clicks, resolving conflicts based on euclidean distance to labeled superpixels. Sun et al. [60] employ a two-stage approach for boundary prediction: a CNN predicts the initial contour, followed by a GCN trained for post-processing to predict vertex offsets from the ground-truth boundary.

C. 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.

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. 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. Längkvist et al. [19] propose a real-time annotation tool for CT scans, training CNNs from scratch with human interactions as labels. They explore the efficiency-accuracy trade-off, comparing small and large models in online learning. ECONet [65] employs a small CNN model with a single convolution layer. Scribbles from the user are utilized directly as ground-truth masks, with fixed-size patches around each voxel used for model updates. Sliding window inference is then applied to the entire volume for Coronavirus Disease 2019 (COVID-19) lung lesion segmentation. Asad et al. [85] enhance ECONet [65] with an adaptive loss that spreads scribble influence to neighboring similar regions. They further prune uncertain predictions below a confidence threshold during weight updates.

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. BIFSeg [6] employs a user-provided bounding box for initial segmentation, followed by correction using scribbles. The model fine-tunes with a weighted loss function based on these scribbles. Dhara et al. [8] expand the fine-tuning step into an iterative loop, where the annotator corrects model predictions with scribble-based GraphCut [131]. These corrections are then used to update a CNN model in real-time. Chao et al. [18] propagate user corrections to neighboring slices by updating the model with a distance-based loss function. Boers et al. [21] use a loss function that assigns higher weights to missegmented voxels from user-provided scribbles, while other voxels are weighted based on their distance to the scribbles. Sambaturu et al. [41] propose an efficient model-agnostic fine-tuning scheme using user-based scribbles. They dilate the scribbles with region growing and introduce an L2 regularization term for weight updates, ensuring stability in the model's predictions.

V. 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.

A. Segmentation Targets, Imaging Modalities, and Evaluation Metrics

1) Segmentation Targets: We distinguish segmentation targets in two primary categories: 1) anatomical structures and cells; and 2) pathologies. The categorization depends on whether a method's primary focus is on specific anatomical structures, distinct pathologies, or both (noted in $n = 7$ of all 121 studies). The number of methods per specific anatomy or pathology is depicted in Fig. 6. Prominent anatomical regions encompass the brain, prostate, and cardiac structures as well as abdominal organs featuring the liver, spleen, kidney, pancreas, stomach, and gallbladder. Thoracic organs are less prominent, including lungs, aorta, esophagus, and cardiac structures, and whole-body structures like bones and blood vessels. Further notable regions of interest which are combined in the "Other"-category encompass lymph nodes, the Z-line, spine, cartilage, and skin. Furthermore, techniques using microscopy and OCT data are predominantly geared towards cell segmentation, targeting blood cells, testicular cells, neurons, or cell nuclei.

Pathological targets exhibit notably less diversity compared to their anatomical counterparts. The prevalent pathologies tend to concentrate primarily within the brain ($n = 21$) and liver regions ($n = 12$), largely owing to the prominence of datasets like: 1) BraTS [162] for brain cancer; 2) as well as MSD [163] and LiTS [164] for liver cancer. Beyond brain and liver cancer, a few specific targets emerge as representative of certain imaging modalities. Notably, COVID-19 lung lesions stand out in X-Rays, while skin lesions take precedence in dermoscopy. Colon cancer and polyps serve as typical examples in colonoscopy imaging. Other relevant pathologies encompass lung, breast, kidney, and thyroid cancer. The "Other"-category consists of less frequently encountered targets such as head and neck cancer, cervical, pancreatic, prostatic, and esophageal cancer, hematomas, and foot ulcers.

2) Imaging Modalities: Radiological modalities, particularly CT ($n = 65$) and MRI ($n = 42$), dominate the imaging modalities and are featured in the most reviewed methods. This prevalence can be attributed to the existence of popular public datasets from segmentation challenge competitions like MSD [163] and BraTS [162]. These challenges frequently release their training data publicly, incentivizing the adoption of these imaging modalities in many approaches. Subsequent to CT and MRI, ultrasound is the choice for $n = 18$ out of 121 approaches, frequently applied in cardiac imaging, mammography,

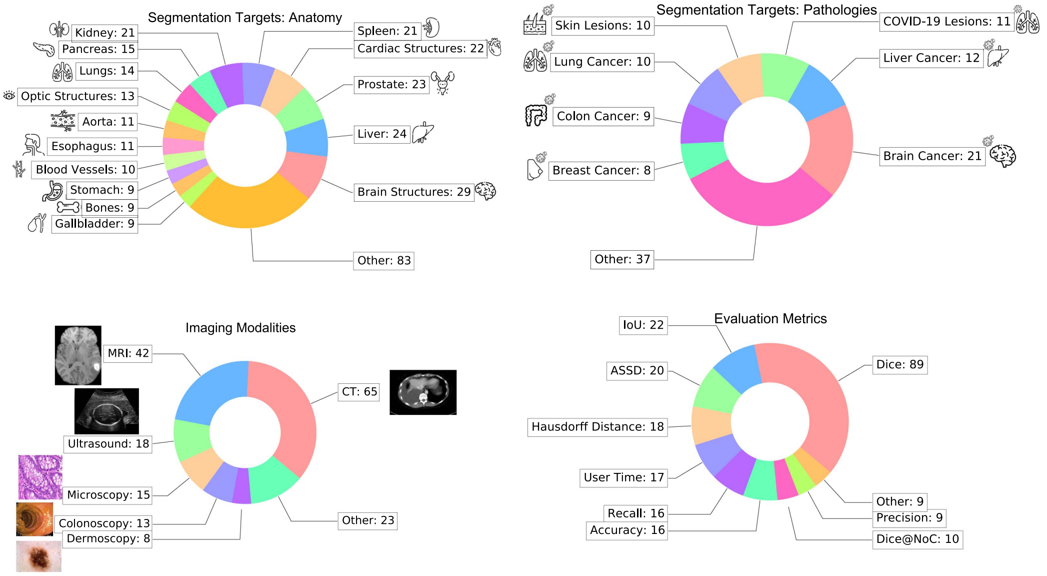


Fig. 6. Distribution of segmentation targets in the anatomy (top left) and in the pathology (top right), imaging modalities (bottom left), and evaluation metrics (bottom right) among all reviewed papers. The numbers represent the number of papers in that category. The icons on the top row are designed by Flaticon.com.

or fetal ultrasound. Microscopy finds application in n = 15 out of 121 reviewed methods, predominantly in pathology for tumor or cancer cell identification. Colonoscopy stands as an imaging modality exclusively dedicated to polyp and/or colon cancer segmentation. Dermoscopy, on the other hand, specializes in skin lesion segmentation. Less frequently encountered imaging modalities in interactive models encompass OCT, X-Ray, fundus imaging, and PET/CT. For a comprehensive listing of segmentation targets and imaging modalities utilized by each reviewed method, refer to Tables VIII and IX in the Appendixes.

3) Evaluation Metrics: An adequate selection of evaluation metrics is crucial for meaningful assessment of segmentation methods, and thus, for trustworthy deployment in practice as well as scientific progress of the field. A large-scale investigation recently found that current medical image segmentation is subject to a substantial extent of pitfalls related to evaluation metrics [170]. The study reveals various shortcomings of the popular Dice Similarity Coefficient (DSC). At the same time, a follow-up study termed “Metrics Reloaded” provides a standardized framework for avoiding these pitfalls and selecting adequate metrics for a given problem [165]. One major finding was that performance should always be assessed by multiple metrics to account for failure modes such as of the DSC. Fig. 6 depicts the evaluation metrics employed by the reviewed studies. As expected, the most-used metric is DSC (n = 89). However, oftentimes the DSC is the only reported metric for segmentation performance (n = 29). Another common problem is the reporting of redundant metrics, such as reporting both DSC and Intersection over Union (n = 19). Remarkably, despite its widespread use as a complementary metric to DSC in non-interactive segmentation and its endorsement by “Metrics Reloaded” for various settings, the Normalized Surface Distance [171] appears in only 2 out of 121 studies reviewed.

The incorporation of user-centered metrics is crucial for devising user-friendly and intuitive methods, particularly in the context of human-in-the-loop approaches. However, there is a noticeable scarcity of user-centric metrics in the reviewed studies. Some studies report the User Time (n = 17), quantifying the active annotator’s labeling time in seconds, or the Dice@NoC (n = 10), measuring the Dice score at a predefined Number of Clicks (NoC). Furthermore, the “Other”-category includes usability metrics like NASA-TLX [147] and the System Usability Scale [148], although these are seldom utilized.

B. *Emergence of Foundation Models*

In early 2023, the Segment Anything Model (SAM) [137] emerged, introducing an approach that involves large-scale training on over 1 billion segmentation masks. Although SAM’s initial training dataset (SA-1B) primarily comprises 2D natural images, several works have showcased its adaptability to medical data, spanning both 2D (such as dermoscopy and fundus) and 3D imaging modalities (including CT, MRI, and PET/CT). This versatility is achieved through targeted fine-tuning on medical data [120]. In the case of 3D images, it commonly involves using 2D axial slices [120] or integration of specialized 2D-to-3D adapters into the model [118].

SAM has shown a good generalization on multiple imaging modalities and tasks, especially on 2D modalities [120] utilizing its bounding box prompting capability. This light-weight adaptability has caused an unprecedented acceleration in the field of deep interactive medical image segmentation as evidenced by 29 proposed medical SAM-adaptations in only a few months at the time of writing. Thus, SAM has demonstrated the potential of utilizing foundation models for medical interactive segmentation. Further, due to its generalization and zero-shot capabilities, it seems to foster a trend towards evaluating methods on a larger number of tasks as some SAM-based approaches are evaluated on over 30 public medical datasets [119], [120].

C. Reproducibility and Availability

In recent years, the field of interactive medical segmentation has witnessed a surge in the emergence of new approaches. There is a promising shift towards enhanced reproducibility, with an increasing number of research papers releasing their code, often accompanied by detailed instructions for replicating results, and in some instances, providing pre-trained model weights. This openness and transparency in sharing resources are further bolstered by the presence of open-source projects like MONAI Label [149], AnatomySketch [70], RIL-Contour [150], BioMedisa [151], MITK [173], and PyMIC [174] which greatly facilitate the development and deployment of interactive deep medical models. Non-deep learning projects such as ilastik [168], ITK-Snap [169], and Li et al. [172], have also contributed to the open source development of interactive models and are widely used in the community. Additionally, this positive trajectory benefits from a growing reliance on openly available challenge datasets sourced from platforms such as Kaggle (www.kaggle.com), Grand Challenge (www.grand-challenge.org), and Synapse (www.synapse.org), promoting collaborative research and advancing the state of the art in the domain. All these tendencies are illustrated in Fig. 1. In our Appendixes, we provide links to code repositories of all reviewed studies with publicly available code as well as links to all 185 public datasets used by the reviewed methods, streamlining access for future researchers.

D. Comparison Graph

Finally, we investigate the field's practice of comparing proposed methods against relevant baselines. Since the scientific merit of a proposed method is measured as the gain over existing solutions, a comprehensive and up-to-date set of baseline methods is crucial for scientific progress in the field. Fig. 7 gives an overview over the comparison practices in deep interactive segmentation of medical images.

The most remarkable observation is the fact that a large fraction ($n = 46$) of the 121 reviewed studies *do not compare against any prior work*. Another portion compares exclusively against "classical methods" ($n = 6$), i.e., non-deep learning-based methods proposed before 2016, or exclusively against DeepIGeoS [5] ($n = 3$). Additionally, a large portion of studies ($n = 37$) compare only to interactive methods which are not trained on medical data, such as DIOS [140], Polygon-RNN [175], DEXTR [176], Latent Diversity (LD) [179], BRS [178], f-BRS [177], and SAM [137]. Despite their shared characteristics, even methods from the same node of the presented taxonomy tree (see circle color in Fig. 7) are most often not compared against each other. Finally, the described acceleration of the field caused by the introduction of SAM seems to compromise the rigor of evaluation given that none of these approaches compares to methods other than the original non-medical SAM. This overall concerning status of a severe lack of cross-comparison in the field comes as a surprise given the positive trends towards reproducibility shown in Fig. 1.

VI. Discussion and Future Directions

Based on the key trends we have identified in Section V, we now derive and discuss the major challenges and opportunities for the field of deep interactive medical image segmentation. The discussion aims to provide a succinct summary of the field's current trajectory while simultaneously identifying pivotal areas where course corrections are necessary.

A. Positive Trends

1) Momentum in Research and Adaptation: The increasing number of publications each year reflects significant momentum and rapid advancements in the field. Additionally, the fast adoption of new paradigms, such as SAM [137], exemplifies the field's dynamic and responsive nature to emerging concepts and technologies.

2) Enhanced Reproducibility and Open-Source Engagement: There has been a notable surge in the use of open-source methods and public datasets. This trend not only facilitates more accessible development of customized models but also encourages the sharing of these models within the community. The proliferation of open-source frameworks specifically designed for interactive segmentation, like MONAI Label [149] and AnatomySketch [70], further underlines this commitment to reproducibility and collaborative growth.

B. Challenges and Opportunities

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.

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].

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

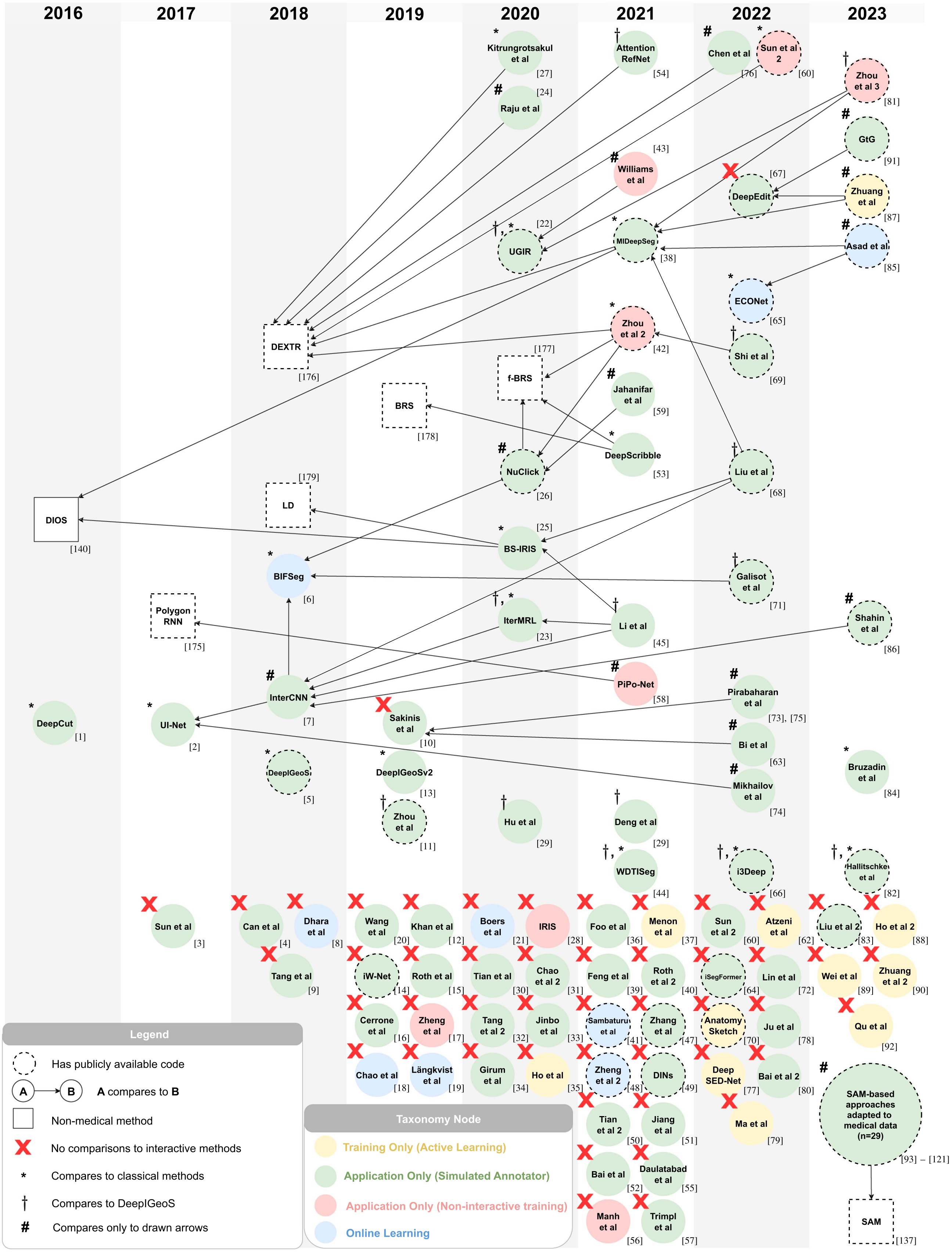


Fig. 7. Comparison graph of all the reviewed methods. Nodes are ordered by initial submission year left to right. Classical methods denote non-deep learning-based interactive methods before 2016. The star (∗) and the dagger (†) are introduced to reduce the visual load in the figure caused by too many arrows.

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.

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 VI-B2 and to report the performance before the first interaction to assess how the task is addressed with non-interactive methods.

C. Text-Based Interactive Segmentation

Text-based interactions are a promising future direction for interactive methods, gaining traction by using text queries to specify segmentation targets. Recent methods use text in two ways: (1) *directly* as an additional input to the interactive model; (2) *indirectly* via visual grounding, where text is used to produce a bounding box, which the interactive model uses as an interaction cue. Direct text-based models [183], [184], [185] leverage image-text pairs for contrastive pre-training, aligning text and image embeddings of the same targets. Indirect text-based models [181], [182] use text as input to a detection model, utilizing the resulting bounding box for interaction. Text-based interactions are expanding the boundaries of the interactive field by exploring novel interactions beyond imaging, extending into the speech domain and other modalities.

VII. Conclusion

In conclusion, our systematic review and the accompanying taxonomy tree stand as a pivotal resource for both researchers and practitioners within the field of deep interactive medical image segmentation. For researchers, this work simplifies the task of locating pertinent related studies, thereby enhancing the quality and relevance of their methodological proposals and evaluations. Practitioners, meanwhile, are empowered to swiftly identify and select methods that are optimally suited to their unique problem scenarios. Additionally, our review has not only identified key trends within the field but also thoroughly discussed the related challenges and opportunities for the future. Most importantly, we have pinpointed a concerning lack of scientific rigor in the evaluation of methods. This critical insight underlines the need for more standardized and systematic benchmarking practices in the field. Overall, we believe this work represents an important step towards implementing such standardized approaches, thereby fostering the development of more reliable, efficient, and effective solutions in deep interactive medical image segmentation.

Acknowledgment

The present contribution is supported by the Helmholtz Association under the Joint Research School "HIDSS4Health – Helmholtz Information and Data Science School for Health."

References

[1] M. Rajchl et al., "DeepCut: Object segmentation from bounding box annotations using convolutional neural networks," *IEEE Trans. Med. Imag.*, vol. 36, no. 2, pp. 674–683, Feb. 2017.

[2] M. Amrehn et al., "UI-net: Interactive artificial neural networks for iterative image segmentation based on a user model," in *Proc. Eurographics Workshop Vis. Comput. Biol. Med.*, 2017, pp. 143–147.

[3] J. Sun, Y. Shi, Y. Gao, and D. Shen, "A point says a lot: An interactive segmentation method for MR prostate via one-point labeling," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2017, pp. 220–228.

[4] Y. B. Can et al., "Learning to segment medical images with scribble-supervision alone," in *Proc. Int. Workshop Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*, 2018, pp. 236–244.

[5] G. Wang et al., "DeepIGeoS: A deep interactive geodesic framework for medical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1559–1572, Jul. 2018.

[6] G. Wang et al., "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Trans. Med. Imag.*, vol. 37, no. 7, pp. 1562–1573, Jul. 2018.

[7] G. Bredell, C. Tanner, and E. Konukoglu, "Iterative interaction training for segmentation editing networks," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2018, pp. 363–370.

[8] A. K. Dhara et al., "Segmentation of post-operative glioblastoma in MRI by U-Net with patient-specific interactive refinement," in *Proc. Int. MICCAI Brainlesion Workshop*, 2019, pp. 115–122.

[9] Y. Tang, A. P. Harrison, M. Bagheri, J. Xiao, and R. M. Summers, "Semi-automatic RECIST labeling on CT scans with cascaded convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2018, pp. 405–413.

[10] T. Sakinis et al., "Interactive segmentation of medical images through fully convolutional neural networks," 2019, *arXiv: 1903.08205v1*.

[11] B. Zhou, L. Chen, and Z. Wang, "Interactive deep editing framework for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2019, pp. 329–337.

[12] S. Khan, A. H. Shahin, J. Villafruela, J. Shen, and L. Shao, "Extreme points derived confidence map as a cue for class-agnostic interactive segmentation using deep neural network," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2019, pp. 66–73.

[13] W. Lei, H. Wang, R. Gu, S. Zhang, S. Zhang, and G. Wang, "DeepIGeoS-V2: Deep interactive segmentation of multiple organs from head and neck images with lightweight CNNs," in *Proc. LABELS Workshop MICCAI*, 2019, pp. 61–69.

[14] G. Aresta et al., "iW-Net: An automatic and minimalistic interactive lung nodule segmentation deep network," *Sci. Rep.*, vol. 9, no. 1, pp. 1–9, Aug. 2019.

[15] H. Roth et al., "Weakly supervised segmentation from extreme points," in *Proc. LABELS Workshop MICCAI*, 2019, pp. 42–50.

[16] L. Cerrone, A. Zeilmann, and F. A. Hamprecht, "End-to-end learned random walker for seeded image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12559–12568.

[17] H. Zheng, Y. Chen, X. Yue, and C. Ma, "Deep interactive segmentation of uncertain regions with shadowed sets," in *Proc. 3 rd Int. Symp. Image Comput. Digit. Med.*, 2019, pp. 244–248.

[18] C.-H. Chao et al., "Radiotherapy target contouring with convolutional gated graph neural network," 2019, *arXiv: 1904.03086v1*.

[19] M. Längkvist, J. Widell, P. Thunberg, A. Loutfi, and M. Lidén, "Interactive user interface based on convolutional auto-encoders for annotating CT-scans," 2019, *arXiv: 1904.11701v1*.

[20] X. Wang, L. Zhang, H. Roth, D. Xu, and Z. Xu, "Interactive 3D segmentation editing and refinement via gated graph neural networks," in *Proc. Int. Workshop Graph Learn. Med. Imag.*, 2019, pp. 9–17.

[21] T. Boers et al., "Interactive 3D U-net for the segmentation of the pancreas in computed tomography scans," *Phys. Med. Biol.*, vol. 65, no. 6, Mar. 2020, Art. no. 065002.

[22] G. Wang, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, and S. Zhang, "Uncertainty-guided efficient interactive refinement of fetal brain segmentation from stacks of MRI slices," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 279–288.

[23] X. Liao et al., "Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9394–9402.

[24] A. Raju et al., "User-guided domain adaptation for rapid annotation from user interactions: A study on pathological liver segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 457–467.

[25] C. Ma et al., "Boundary-aware supervoxel-level iteratively refined interactive 3D image segmentation with multi-agent reinforcement learning," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2563–2574, Oct. 2021.

[26] N. A. Koohbanani, M. Jahanifar, N. Z. Tajadin, and N. Rajpoot, "NuClick: A deep learning framework for interactive segmentation of microscopic images," *Med. Image Anal.*, vol. 65, 2020, Art. no. 101771.

[27] T. Kitrungrotsakul, I. Yutaro, L. Lin, R. Tong, J. Li, and Y.-W. Chen, "Interactive deep refinement network for medical image segmentation," 2020, *arXiv: 2006.15320v1*.

[28] A. Pepe et al., "IRIS: Interactive real-time feedback image segmentation with deep learning," in *Proc. SPIE Med. Imag. Biomed. Appl. Mol. Struct. Funct. Imag.*, 2020, pp. 181–186.

[29] W. Hu et al., "Error attention interactive segmentation of medical image through matting and fusion," in *Proc. Int. Workshop Mach. Learn. Med. Imag.*, 2020, pp. 11–20.

[30] Z. Tian et al., "Graph-convolutional-network-based interactive prostate segmentation in MR images," *Med. Phys.*, vol. 47, no. 9, pp. 4164–4176, Jun. 2020.

[31] C.-H. Chao, H.-T. Cheng, T.-Y. Ho, L. Lu, and M. Sun, "Interactive radiotherapy target delineation with 3D-fused context propagation," 2020, *arXiv: 2012.06873v1*.

[32] Y. Tang, K. Yan, J. Xiao, and R. M. Summers, "One click lesion RECIST measurement and segmentation on CT scans," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 573–583.

[33] H. Jinbo, T. Kitrungrotsaku, Y. Iwamoto, L. Lin, H. Hu, and Y.-W. Chen, "Development of an interactive semantic medical image segmentation system," in *Proc. IEEE 9th Glob. Conf. Consum. Electron.*, 2020, pp. 678–681.

[34] K. B. Girum, G. Créhange, R. Hussain, and A. Lalande, "Fast interactive medical image segmentation with weakly supervised deep learning method," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, pp. 1437–1444, Jul. 2020.

[35] D. J. Ho et al., "Deep interactive learning: An efficient labeling approach for deep learning-based osteosarcoma treatment response assessment," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2020, pp. 540–549.

[36] M. X. et al., "Interactive segmentation for COVID-19 infection quantification on longitudinal CT scans," *IEEE Access*, vol. 11, pp. 77596–77607, 2023.

[37] A. Menon, P. Singh, P. Vinod, and C. Jawahar, "Interactive learning for assisting whole slide image annotation," in *Proc. Asian Conf. Pattern Recognit.*, 2021, pp. 504–517.

[38] X. Luo et al., "MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning," *Med. Image Anal.*, vol. 72, Aug. 2021, Art. no. 102102.

[39] R. Feng et al., "Interactive few-shot learning: Limited supervision, better medical image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 10, pp. 2575–2588, Oct. 2021.

[40] H. R. Roth, D. Yang, Z. Xu, X. Wang, and D. Xu, "Going to extremes: Weakly supervised medical image segmentation," *Mach. Learn. Knowl. Extraction*, vol. 3, no. 2, pp. 507–524, Jun. 2021.

[41] B. Sambaturu, A. Gupta, C. Jawahar, and C. Arora, "Efficient and generic interactive segmentation framework to correct mispredictions during clinical evaluation of medical images," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 625–635.

[42] T. Zhou, L. Li, G. Bredell, J. Li, and E. Konukoglu, "Quality-aware memory network for interactive volumetric image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 560–570.

[43] H. Williams et al., "Interactive segmentation via deep learning and B-spline explicit active surfaces," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2021, pp. 315–325.

[44] X. Li et al., "WDTISeg: One-stage interactive segmentation for breast ultrasound image using weighted distance transform and shape-aware compound loss," *Appl. Sci.*, vol. 11, no. 14, Jul. 2021, Art. no. 6279.

[45] W. Li et al., "Interactive medical image segmentation with self-adaptive confidence calibration," *Front. Inf. Technol. Electron. Eng.*, vol. 24, no. 9, pp. 1332–1348, Sep. 2023.

[46] J. Deng and X. Xie, "3D interactive segmentation with semi-implicit representation and active learning," *IEEE Trans. Image Process.*, vol. 30, pp. 9402–9417, 2021.

[47] J. Zhang et al., "Interactive medical image segmentation via a point-based interaction," *Artif. Intell. Med.*, vol. 111, Jan. 2021, Art. no. 101998.

[48] E. Zheng, Q. Yu, R. Li, P. Shi, and A. Haake, "A continual learning framework for uncertainty-aware interactive image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 6030–6038.

[49] J.-W. Zhang et al., "DINs: Deep interactive networks for neurofibroma segmentation in neurofibromatosis type 1 on whole-body MRI," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 2, pp. 786–797, Feb. 2022.

[50] Z. Tian et al., "Interactive prostate mr image segmentation based on ConvLSTMs and GGNN," *Neurocomputing*, vol. 438, pp. 84–93, May 2021.

[51] D. Jiang et al., "Residual refinement for interactive skin lesion segmentation," *J. Biomed. Semantics*, vol. 12, no. 1, 2021, Art. no. 22.

[52] Y. Bai, G. Sun, Y. Li, L. Shen, and L. Zhang, "Progressive medical image annotation with convolutional neural network-based interactive segmentation method," in *Proc. SPIE Med. Imag. Image Process.*, 2021, pp. 732–742.

[53] S. Cho, H. Jang, J. W. Tan, and W.-K. Jeong, "DeepScribble: Interactive pathology image segmentation using deep neural networks with scribbles," in *Proc. IEEE 18th Int. Symp. Biomed. Imag.*, 2021, pp. 761–765.

[54] T. Kitrungrotsakul et al., "Attention-RefNet: Interactive attention refinement network for infected area segmentation of COVID-19," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 7, pp. 2363–2373, Jul. 2021.

[55] R. Daulatabad, R. Vega, J. L. Jaremko, J. Kapur, A. R. Hareendranathan, and K. Punithakumar, "Integrating user-input into deep convolutional neural networks for thyroid nodule segmentation," in *Proc. 43 rd Annu. Int. Conf IEEE Eng. Med. Biol. Soc.*, 2021, pp. 2637–2640.

[56] X. H. Manh et al., "Interactive Z-line segmentation tool for upper gastrointestinal endoscopy images using binary partition tree and U-Net," in *Proc. Int. Conf. Comput. Commun. Technol. Res. Innov. Vis. Future*, 2021, pp. 1–6.

[57] M. J. Trimpl, D. Boukerroui, E. P. Stride, K. A. Vallis, and M. J. Gooding, "Interactive contouring through contextual deep learning," *Med. Phys.*, vol. 48, no. 6, pp. 2951–2959, Mar. 2021.

[58] Y. Fang, D. Zhu, N. Zhou, L. Liu, and J. Yao, "PiPo-Net: A semi-automatic and polygon-based annotation method for pathological images," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 2978–2984.

[59] M. Jahanifar, N. Z. Tajeddin, N. A. Koohbanani, and N. M. Rajpoot, "Robust interactive semantic segmentation of pathology images with minimal user input," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 674–683.

[60] L. Sun, Z. Tian, Z. Chen, W. Luo, and S. Du, "An efficient interactive segmentation framework for medical images without pre-training," *Med. Phys.*, vol. 50, no. 4, pp. 2239–2248, Apr. 2023.

[61] M. Shahedi, J. D. Dormer, M. Halicek, and B. Fei, "The effect of image annotation with minimal manual interaction for semiautomatic prostate segmentation in CT images using fully convolutional neural networks," *Med. Phys.*, vol. 49, no. 2, pp. 1153–1160, Feb. 2022.

[62] A. Atzeni et al., "Deep active learning for suggestive segmentation of biomedical image stacks via optimisation of dice scores and traced boundary length," *Med. Image Anal.*, vol. 81, Oct. 2022, Art. no. 102549.

[63] L. Bi, M. Fulham, and J. Kim, "Hyper-fusion network for semi-automatic segmentation of skin lesions," *Med. Image Anal.*, vol. 76, 2022, Art. no. 102334.

[64] Q. Liu, Z. Xu, Y. Jiao, and M. Niethammer, "iSegFormer: Interactive segmentation via transformers with application to 3D knee MR images," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2022, pp. 464–474.

[65] M. Asad, L. Fidon, and T. Vercauteren, "ECONet: Efficient convolutional online likelihood network for scribble-based interactive segmentation," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2022, pp. 35–47.

[66] K. Gotkowski, C. Gonzalez, I. Kaltenborn, R. Fischbach, A. Bucher, and A. Mukhopadhyay, "i3Deep: Efficient 3D interactive segmentation with the nnU-Net," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2022, pp. 441–456.

[67] A. Diaz-Pinto et al., "DeepEdit: Deep editable learning for interactive segmentation of 3D medical images," in *Proc. DALI Workshop MICCAI*, 2022, pp. 11–21.

[68] W. Liu, C. Ma, Y. Yang, W. Xie, and Y. Zhang, "Transforming the interactive segmentation for medical imaging," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2022, pp. 704–713.

[69] L. Shi, X. Zhang, Y. Liu, and X. Han, "A hybrid propagation network for interactive volumetric image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2022, pp. 673–682.

[70] M. Zhuang et al., "AnatomySketch: An extensible open-source software platform for medical image analysis algorithm development," *J. Digit. Imag.*, vol. 35, no. 6, pp. 1623–1633, Jun. 2022.

[71] G. Galisot, J.-Y. Ramel, T. Brouard, E. Chaillou, and B. Serres, "Visual and structural feature combination in an interactive machine learning system for medical image segmentation," *Mach. Learn. Appl.*, vol. 8, Jun. 2022, Art. no. 100294.

[72] Z. Lin, Z. Zhang, L.-H. Han, and S.-P. Lu, "Multi-mode interactive image segmentation," in *Proc. 30th ACM Int. Conf. Multimedia*, 2022, pp. 905–914.

[73] R. Pirabaharan and N. Khan, "Interactive segmentation using U-Net with weight map and dynamic user interactions," in *Proc. 43 rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2022, pp. 4754–4757.

[74] I. Mikhailov, B. Chauveau, N. Bourdel, and A. Bartoli, "A deep learning-based interactive medical image segmentation framework," in *Proc. AMAI Workshop MICCAI*, 2022, pp. 98–107.

[75] R. Pirabaharan and N. Khan, "Improving interactive segmentation using a novel weighted loss function with an adaptive click size and two-stream fusion," in *Proc. IEEE 8th Int. Conf. Multimedia Big Data*, 2022, pp. 7–12.

[76] X. Chen et al., "Balancing regional and global information: An interactive segmentation framework for ultrasound breast lesion," *Biomed. Signal Process. Control*, vol. 77, Aug. 2022, Art. no. 103723.

[77] S. Liang et al., "Deep SED-Net with interactive learning for multiple testicular cell types segmentation and cell composition analysis in mouse seminiferous tubules," *Cytometry Part A*, vol. 101, no. 8, pp. 658–674, Apr. 2022.

[78] M. Ju, M. Lee, J. Lee, J. Yang, S. Yoon, and Y. Kim, "All you need is a few dots to label CT images for organ segmentation," *Appl. Sci.*, vol. 12, no. 3, Jan. 2022, Art. no. 1328.

[79] W. Ma, S. Zheng, L. Zhang, H. Zhang, and Q. 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 *Proc. IEEE 19th Int. Symp. Biomed. Imag.*, 2022, pp. 1–5.

[80] T. Bai et al., "A proof-of-concept study of artificial intelligence–assisted contour editing," *Radiol. Artif. Intell.*, vol. 4, no. 5, Sep. 2022, Art. no. e210214.

[81] T. Zhou, L. Li, G. Bredell, J. Li, J. Unkelbach, and E. Konukoglu, "Volumetric memory network for interactive medical image segmentation," *Med. Image Anal.*, vol. 83, Jan. 2023, Art. no. 102599.

[82] V. J. Hallitschke et al., "Multimodal interactive lung lesion segmentation: A framework for annotating PET/CT images based on physiological and anatomical cues," in *Proc. IEEE 18th Int. Symp. Biomed. Imag.*, 2023, pp. 1–5.

[83] Q. Liu et al., "Exploring cycle consistency learning in interactive volume segmentation," 2023, *arXiv:2303.06493v2*.

[84] A. Bruzadin, M. Boaventura, M. Colnago, R. G. Negri, and W. Casaca, "Learning label diffusion maps for semi-automatic segmentation of lung CT images with COVID-19," *Neurocomputing*, vol. 522, pp. 24–38, Feb. 2023.

[85] M. Asad et al., "Adaptive multi-scale online likelihood network for AI-assisted interactive segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2023, pp. 564–574.

[86] A. H. Shahin, Y. Zhuang, and N. El-Zehiry, "From sparse to precise: A practical editing approach for intracardiac echocardiography segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2023, pp. 766–775.

[87] M. Zhuang et al., "Efficient contour-based annotation by iterative deep learning for organ segmentation from volumetric medical images," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 18, no. 2, pp. 379–394, Sep. 2022.

[88] D. J. Ho et al., "Deep interactive learning-based ovarian cancer segmentation of H&E-stained whole slide images to study morphological patterns of BRCA mutation," *J. Pathol. Inform.*, vol. 14, Jan. 2023, Art. no. 100160.

[89] Z. Wei, J. Ren, S. S. Korreman, and J. Nijkamp, "Towards interactive deep-learning for tumour segmentation in head and neck cancer radiotherapy," *Phys. Imag. Radiat. Oncol.*, vol. 25, Jan. 2023, Art. no. 100408.

[90] M. Zhuang, Z. Chen, Y. Yang, L. Kettunen, and H. Wang, "Annotation-efficient training of medical image segmentation network based on scribble guidance in difficult areas," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 19, no. 1, pp. 87–96, May 2023.

[91] Z. Marinov, R. Stiefelhagen, and J. Kleesiek, "Guiding the Guidance: A comparative analysis of user guidance signals for interactive segmentation of volumetric images," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Interv.*, 2023, pp. 637–647.

[92] C. Qu et al., "AbdomenAtlas-8 K: Annotating 8,000 CT volumes for multi-organ segmentation in three weeks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 36620–36636.

[93] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang, "Segment anything model for medical image analysis: An experimental study," *Med. Image Anal.*, vol. 89, Oct. 2023, Art. no. 102918.

[94] R. Deng et al., "Segment anything model (SAM) for digital pathology: Assess zero-shot segmentation on whole slide imaging," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2023. [Online]. Available: https://openreview.net/group?id=MIDL.io/2023/Conference

[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," 2023, *arXiv:2304.04738v3*.

[96] F. Putz et al., "The segment anything foundation model achieves favorable brain tumor autosegmentation accuracy on MRI to support radiotherapy treatment planning," 2023, *arXiv:2304.07875v1*.

[97] C. Hu and X. Li, "When SAM meets medical images: An investigation of segment anything model (SAM) on multi-phase liver tumor segmentation," 2023, *arXiv:2304.08506v6*.

[98] T. Chen et al., "SAM fails to segment anything?–SAM-adapter: Adapting SAM in underperformed scenes: Camouflage, shadow, and more," 2023, *arXiv:2304.09148v3*.

[99] J. Wu et al., "Medical SAM adapter: Adapting segment anything model for medical image segmentation," 2023, *arXiv:2304.12620v7*.

[100] Z. Qiu, Y. Hu, H. Li, and J. Liu, "Learnable ophthalmology SAM," 2023, *arXiv:2304.13425v1*.

[101] S. He, R. Bao, J. Li, P. E. Grant, and Y. Ou, "Accuracy of segment-anything model (SAM) in medical image segmentation tasks," 2023, *arXiv:2304.09324v3*.

[102] P. Shi et al., "Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation," *Diagnostics*, vol. 13, no. 11, Jun. 2023, Art. no. 1947.

[103] B. Wang, A. Aboah, Z. Zhang, and U. Bagci, "GazeSAM: What you see is what you segment," 2023, *arXiv:2304.13844v1*.

[104] M. Hu, Y. Li, and X. Yang, "SkinSAM: Empowering skin cancer segmentation with segment anything model," 2023, *arXiv:2304.13973v1*.

[105] A. Wang, M. Islam, M. Xu, Y. Zhang, and H. Ren, "SAM meets robotic surgery: An empirical study in robustness perspective," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv. Workshops*, 2023, pp. 234–244.

[106] D. Cheng, Z. Qin, Z. Jiang, S. Zhang, Q. Lao, and K. Li, "SAM on medical images: A. comprehensive study on three prompt modes," 2023, *arXiv:2305.00035v1*.

[107] C. Mattjie et al., "Exploring the zero-shot capabilities of the segment anything model (SAM) in 2D medical imaging: A comprehensive evaluation and practical guideline," 2023, *arXiv:2305.00109v2*.

[108] Y. Li, M. Hu, and X. Yang, "Polyp-SAM: Transfer SAM for polyp segmentation," 2023, *arXiv:2305.00293v1*.

[109] J. Wu, J. Zhu, Y. Liu, Y. Jin, and M. Xu, "One-prompt to segment all medical images," 2023, *arXiv:2305.10300v3*.

[110] M. Hu, Y. Li, and X. Yang, "BreastSAM: A study of segment anything model for breast tumor detection in ultrasound images," 2023, *arXiv:2305.12447v1*.

[111] D. Lee, J. Park, S. Cook, S.-J. D. YooLee, and H. Choi, "IAMSAM: Image-based analysis of molecular signatures using the segment-anything model," 2023, *bioRxiv:2023.05.25.542052v1*.

[112] Y. Gao, W. Xia, D. Hu, and X. Gao, "DeSAM: Decoupling segment anything model for generalizable medical image segmentation," 2023, *arXiv:2306.00499v1*.

[113] C. Shen, W. Li, Y. Zhang, and X. Wang, "Temporally-extended prompts optimization for SAM in interactive medical image segmentation," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2023, pp. 3550–3557.

[114] G. Ning, H. Liang, Z. Jiang, H. Zhang, and H. Liao, "The potential of 'segment anything'(SAM) for universal intelligent ultrasound image guidance," *BioSci. Trends*, vol. 17, no. 3, pp. 230–233, Mar. 2023.

[115] L. Zhang et al., "Segment anything model (SAM) for radiation oncology," 2023, *arXiv:2306.11730v2*.

[116] W. Lei, X. Wei, X. Zhang, K. Li, and S. Zhang, "MedLSAM: Localize and segment anything model for 3D medical images," 2023, *arXiv:2306.14752v3*.

[117] G. Deng et al., "SAM-U: Multi-box prompts triggered uncertainty estimation for reliable SAM in medical image," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv. Workshops*, 2023, pp. 368–377.

[118] S. Gong et al., "3DSAM-adapter: Holistic adaptation of SAM from 2D to 3D for promptable medical image segmentation," 2023, *arXiv:2306.13465v1*.

[119] Y. Huang et al., "Segment anything model for medical images?," *Med. Image Anal.*, vol. 92, Feb. 2024, Art. no. 103061.

[120] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Commun.*, vol. 15, no. 1, Jan. 2024, Art. no. 654.

[121] S. Roy et al., "SAM.MD: Zero-shot medical image segmentation capabilities of the segment anything model," in *Proc. Int. Conf. Med. Imag. Deep Learn.*, 2023. [Online]. Available: https://openreview.net/group?id=MIDL.io/2023/Conference

[122] H. Nickisch, C. Rother, P. Kohli, and C. Rhemann, "Learning an interactive segmentation system," in *Proc. 7th Indian Conf. Comput. Vis. Graph. Image Process.*, 2010, pp. 274–281.

[123] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, Nov. 1986.

[124] W. Pedrycz, "Shadowed sets: Representing and processing fuzzy sets," *IEEE Trans. Syst., Man, Cybern. B., Cybern.*, vol. 28, no. 1, pp. 103–109, Feb. 1998.

[125] D. Barbosa, T. Dietenbeck, J. Schaerer, J. D'hooge, D. Friboulet, and O. Bernard, "B-spline explicit active surfaces: An efficient framework for real-time 3-D region-based segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 241–251, Jan. 2012.

[126] A. Yezzi Jr, A. Tsai, and A. Willsky, "A fully global approach to image segmentation via coupled curve evolution equations," *J. Vis. Commun. Image Representations*, vol. 13, no. 1, pp. 195–216, Mar. 2002.

[127] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2015, pp. 234–241.

[128] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[129] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, Jan. 2018.

[130] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[131] Y. Y. Boykov and M.-P. Jolly, "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, pp. 105–112.

[132] P. Salembier and L. Garrido, "Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval," *IEEE Trans. Image Process.*, vol. 9, no. 4, pp. 561–576, Apr. 2000.

[133] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, May 2012.

[134] M. Jenkinson et al., "BET2: MR-based estimation of brain, skull and scalp surfaces," in *Proc. 11th Annu. Meeting Org. Hum. Brain Mapping*, 2005, Art. no. 167.

[135] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. R. Roth, and D. Xu, "Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images," in *Proc. Int. MICCAI Brainlesion Workshop*, 2021, pp. 272–284.

[136] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, pp. 203–211, Dec. 2020.

[137] A. Kirillov et al., "Segment anything," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 3992–4003.

[138] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Representations*, 2021. [Online]. Available: https://openreview.net/group?id=ICLR.cc/2021/Conference

[139] D. Moher et al., "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *Ann. Intern. Med.*, vol. 151, no. 4, pp. 264–269, Aug. 2009.

[140] N. Xu, B. Price, S. Cohen, J. Yang, and T. S. Huang, "Deep interactive object selection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 373–381.

[141] C. Rother, V. Kolmogorov, and A. Blake, "GrabCut-interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[142] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 724–732.

[143] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, Sep. 2009.

[144] B. Hariharan, P. Arbeláez, L. Bourdev, S. Maji, and J. Malik, "Semantic contours from inverse detectors," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 991–998.

[145] K. McGuinness and N. E. O'connor, "A comparative evaluation of interactive segmentation algorithms," *Pattern Recognit.*, vol. 43, no. 2, pp. 434–444, Feb. 2010.

[146] M. Eisenmann et al., "Why is the winner the best?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 19955–19966.

[147] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (task load index): Results of empirical and theoretical research," *Adv. Psychol.*, vol. 52, pp. 139–183, Jan. 1988.

[148] J. Brooke, "SUS-a quick and dirty usability scale," *Usability Eval. Ind.*, vol. 189, no. 194, pp. 4–7, 1996.

[149] A. Diaz-Pinto et al., "MONAI Label: A framework for AI-assisted interactive labeling of 3D medical images," 2022, *arXiv:2203.12362v2*.

[150] K. A. Philbrick et al., "RIL-contour: A medical imaging dataset annotation tool for and with deep learning," *J. Digit. Imag.*, vol. 32, pp. 571–581, May 2019.

[151] P. D. Lösel et al., "Introducing Biomedisa as an open-source online platform for biomedical image segmentation," *Nature Commun.*, vol. 11, no. 1, 2020, Art. no. 5577.

[152] S. Mahadevan, P. Voigtlaender, and B. Leibe, "Iteratively trained interactive segmentation," in *Proc. Brit. Mach. Vis. Conf.*, 2018. [Online]. Available: http://bmvc2018.org/programme/BMVC2018.zip

[153] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *Int. J. Comput. Vis.*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[154] S. Gatidis et al., "The autoPET challenge: Towards fully automated lesion segmentation in oncologic PET/CT imaging," 14 Jun. 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2572595/v1]

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

[156] S. D. Olabarriaga and A. W. Smeulders, "Interaction in the segmentation of medical images: A survey," *Med. Image Anal.*, vol. 5, no. 2, pp. 127–142, Jun. 2001.

[157] H. Ramadan, C. Lachqar, and H. Tairi, "A survey of recent interactive image segmentation methods," *Comput. Vis. Media*, vol. 6, pp. 355–384, Aug. 2020.

[158] Ç. Kaymak and A. Uçar, "A brief survey and an application of semantic image segmentation for autonomous driving," in *Handbook of Deep Learning Applications.*, Berlin, Germany: Springer, 2019, pp. 161–200.

[159] D. Tabernik, S. Šela, J. Skvarč, and D. Skočaj, "Segmentation-based deep-learning approach for surface-defect detection," *J. Intell. Manuf.*, vol. 31, no. 3, pp. 759–776, May 2019.

[160] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

[161] M. Bakator and D. Radosav, "Deep learning and medical diagnosis: A review of literature," *Multimodal Technol. Interact.*, vol. 2, no. 3, Aug. 2018, Art. no. 47.

[162] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[163] M. Antonelli et al., "The medical segmentation Decathlon," *Nature Commun.*, vol. 13, no. 1, Jul. 2022, Art. no. 4128.

[164] P. Bilic et al., "The liver tumor segmentation benchmark (LiTS)," *Med. Image Anal.*, vol. 84, Feb. 2023, Art. no. 102680.

[165] L. Maier-Hein et al., "Metrics reloaded: Recommendations for image analysis validation," *Nature Methods*, vol. 21, no. 2, pp. 195–212, Feb. 2024.

[166] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with Gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

[167] A. E. Lefohn, J. E. Cates, and R. T. Whitaker, "Interactive, GPU-based level sets for 3D segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Interv.*, 2003, pp. 564–572.

[168] C. Sommer, C. Straehle, U. Koethe, and F. A. Hamprecht, "Ilastik: Interactive learning and segmentation toolkit," in *Proc. IEEE 8th Int. Symp. Biomed. Imag.*, 2011, pp. 230–233.

[169] P. A. Yushkevich, Y. Gao, and G. Gerig, "ITK-SNAP: An interactive tool for semi-automatic segmentation of multi-modality biomedical images," in *Proc. 38th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2016, pp. 3342–3345.

[170] A. Reinke et al., "Understanding metric-related pitfalls in image analysis validation," *Nature Methods*, vol. 21, no. 2, pp. 182–194, Feb. 2024.

[171] S. Nikolov et al., "Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study," *J. Med. Internet Res.*, vol. 23, no. 7, Jul. 2021, Art. no. e26151.

[172] R. Li and X. Chen, "An efficient interactive multi-label segmentation tool for 2D and 3D medical images using fully connected conditional random field," *Comput. Methods Programs Biomed.*, vol. 213, Jan. 2022, Art. no. 106534.

[173] I. Wolf et al., "The medical imaging interaction toolkit," *Med. Image Anal.*, vol. 9, no. 6, pp. 594–604, Dec. 2005.

[174] G. Wang et al., "PyMIC: A deep learning toolkit for annotation-efficient medical image segmentation," *Comput. Methods Programs Biomed.*, vol. 231, Apr. 2023, Art. no. 107398.

[175] L. Castrejon, K. Kundu, R. Urtasun, and S. Fidler, "Annotating object instances with a Polygon-RNN," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5230–5238.

[176] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool, "Deep extreme cut: From extreme points to object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 616–625.

[177] K. Sofiiuk, I. Petrov, O. Barinova, and A. Konushin, "f-BRS: Rethinking backpropagating refinement for interactive segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8623–8632.

[178] W. -D. Jang and C. -S. Kim, "Interactive image segmentation via backpropagating refinement scheme," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5297–5306.

[179] Z. Li, Q. Chen, and V. Koltun, "Interactive image segmentation with latent diversity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 577–585.

[180] A. S. A. Khaizi, R. A. M. Rosidi, H.-S. Gan, and K. A. Sayuti, "A mini review on the design of interactive tool for medical image segmentation," in *Proc. Int. Conf. Eng. Technol. Technopreneurship*, 2017, pp. 1–5.

[181] R. Biswas, "Polyp-SAM: Can a text guided sam perform better for polyp segmentation?," 2023, *arXiv:2308.06623*.

[182] D. B. Ramesh, R. I. Sridhar, P. Upadhyaya, and R. 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," Authorea Preprints, 2023.

[183] J. Liu et al., "Clip-driven universal model for organ segmentation and tumor detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 21152–21164.

[184] Z. Zhao et al., "One model to rule them all: Towards universal segmentation for medical images with text prompts," 2023, *arXiv:2312.17183*.

[185] T. Koleilat, H. Asgariandehkordi, H. Rivaz, and Y. Xiao, "MedCLIP-SAM: Bridging text and image towards universal medical image segmentation," 2024, *arXiv:2403.20253*.

Zdravko Marinov received the master's degree in computer science from the Karlsruhe Institute of Technology, Germany, in 2021. He is currently working toward the PhD degree in the Computer Vision for Human-Computer Interaction lab with the Karlsruhe Institute of Technology under the Helmholtz Information & Data Science School for Health. His research interests include interactive segmentation using deep learning to accelerate the annotation of medical images and medical image analysis.

Paul F. Jäger is the Principal investigator of the Interactive Machine Learning Group with the German Cancer Research Center (DKFZ) and Helmholtz Imaging. He studied Physics in Karlsruhe, Stockholm, and Melbourne. During his PhD degree at the Karlsruhe Institute of Technology (KIT) and the DKFZ, he spent six months as a Research Intern at Facebook AI Research. Paul's research revolves around building Generalist AI Systems for healthcare with a focus on medical imaging.

Jan Egger has more than 10 years of experience in medical image analysis and AI-guided therapies, including 5 years in postdoctoral R&D. He has published more than 150 peer-reviewed papers, edited several books, and holds PhDs in computer science and human biology, and a Habilitation in computer science. He is currently a professor at the Institute for Artificial Intelligence in Medicine at Essen University Hospital.

Jens Kleesiek received the PhD degree in computer science. He is a full professor with the Institute for Artificial Intelligence in Medicine (IKIM) and the associate director of the West German Cancer Center. He studied medicine in Heidelberg, specializing in radiology and medical informatics, and bioinformatics with the University of Hamburg. His research focuses on self- and weakly-supervised learning to detect clinically relevant patterns and integrate multimodal information for better clinical decision-making.

Rainer Stiefelhagen (Member IEEE) received the Diploma and PhD degrees from Universität Karlsruhe (TH), in 1996 and 2002, respectively. He is a full professor with the Karlsruhe Institute of Technology (KIT), directing the computer vision for Human-Computer Interaction Lab and the Center for Digital Accessibility and Assistive Technology – ACCESS@KIT. His research focuses on image, document, and video analysis for assistive robots, medical image analysis, and technology for visually impaired users.