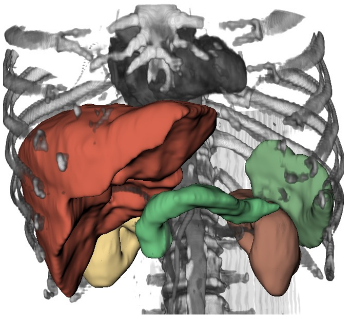
Karlsruher Institut für Technologie

LEVERAGING ANATOMICAL KNOWLEDGE ACROSS THE MODEL DEVELOPMENT LIFECYCLE FOR MEDICAL IMAGE SEGMENTATION

Zur Erlangung des akademischen Grades eines
Doktors der Ingenieurwissenschaften (Dr.-Ing.)

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

Dissertation von
Alexander Jaus



Tag der mündlichen Prüfung: 12.11.2025

Hauptreferent: Prof. Dr.-Ing. Rainer Stiefelhagen
Koreferent: Prof. Dr. med. Dr. rer. nat. Jens Kleesiek

Alexander Jaus: *Leveraging Anatomical Knowledge Across the Model Development Lifecycle for Medical Image Segmentation*

Man is something to be surpassed

— Friedrich Nietzsche

Dedicated to my grandparents Alfred & Maria

ABSTRACT

Image segmentation is a cornerstone of medical image analysis, as it accurately outlines structures of interest in images, thereby connecting the image to a location-specific semantic understanding. While segmentations can be done and are to some extent still done manually, leveraging automated approaches has the potential to significantly support medical personnel, enhance diagnostic accuracy, standardize clinical assessments, and provide robust, objective metrics for monitoring disease progression.

Physicians develop a rich, well-rounded understanding of anatomy through years of study and hands-on experience. Neural networks, on the other hand, derive their representation by contrast from limited, often partly labeled image sets. That broader perspective enables clinicians to spot the subtle twists and turns that distinguish normal anatomy from early signs of disease, a task that remains challenging for neural networks. In this thesis, we aim to narrow that gap from both the perspective of training neural networks and evaluating them. We equip neural networks with a deeper understanding of anatomy and demonstrate how the obtained anatomical knowledge improves pathology segmentation, thereby moving them toward human-level anatomical reasoning. Beyond the training advances, we raise concerns about how progress in pathology segmentation is currently measured, and subsequently develop a new framework that aligns more closely with clinical relevance to quantify model performance.

We build the foundation of this work as a holistic anatomical dataset, curated from a collection of existing but fragmented datasets. We train fragmented models on each of the datasets and subsequently use them to predict their limited knowledge in the form of pseudo labels into a full-body CT (Computed Tomography) dataset, which is used to aggregate all the fragmented knowledge. Through iterative training and anatomically guided refinements, we obtain a high-quality full-body CT dataset that, for the first time, provides labels for the majority of the human body, allowing the training of neural networks capable of understanding large parts of the human anatomy.

Hypothesizing that a better understanding of human anatomy enables networks to segment pathologies and foreign objects more effectively, we investigate how networks equipped with anatomical understanding perform across two tasks: segmenting pathologies in whole-body PET/CT imaging and identifying thoracic abnormalities in chest X-rays. We develop APEx, a novel framework to jointly learn anatomies and pathologies, focusing on their interrelatedness. We find that a rich understanding of the human anatomy benefits the segmentation performance in both imaging domains.

Complementing the advancement in training anatomy-pathology models, we turn towards evaluation. We find that previous approaches, which are largely influenced by semantic segmentation metrics such as Dice, only measure the overlap between ground-truth and predictions. Simply comparing global overlap, however, falls short in terms of the specific characteristics of full-body lesion segmentation. With smaller lesions not being less critical than larger ones, these commonly used metrics fail to address the unique challenges of this field, thereby biasing the model's performance. We address this limitation by developing the Connected-Component (CC) framework, which reweights any standard segmentation metric on a per-component basis, thereby better measuring the models' capabilities to segment tumors across the entire scan, irrespective of the tumor size.

Overall, this thesis bridges the gap between computational and clinical understanding of anatomy by advancing three critical technical pillars: dataset generation, model training, and model evaluation. These contributions firstly provide a holistic anatomical foundation for neural networks, secondly demonstrate the impact of anatomical knowledge on pathology detection, and thirdly provide evaluation metrics that better reflect clinical priorities, thereby advancing medical AI towards a more reliable, anatomically-informed, and clinically meaningful future.

ZUSAMMENFASSUNG

Die Bildsegmentierung ist ein zentraler Bestandteil der medizinischen Bildanalyse, da sie relevante Strukturen in Bildern genau umreißt und so das Bild mit einem positionsspezifischen semantischen Verständnis verbindet. Segmentierungen können und werden zum Teil immer noch von Hand gemacht; der Einsatz automatisierter Ansätze hat jedoch das Potenzial, das medizinische Personal erheblich zu unterstützen, die Diagnosegenauigkeit zu verbessern, klinische Beurteilungen zu standardisieren und robuste, objektive Metriken zur Verlaufsüberwachung von Krankheiten zu berechnen.

Ärzte haben sich in jahrelangem Studium und durch praktische Arbeit ein umfassendes, ganzheitliches Bild der Anatomie gemacht. Neuronale Netze hingegen leiten ihre Repräsentation der Anatomie aus begrenzten, oft nur teilweise gelabelten Bildsätzen ab. Dank dieser breiteren Perspektiven können Kliniker die minimalen Abweichungen zwischen normaler Anatomie und frühen Anzeichen von Krankheiten erkennen, was neuronalen Netzen oft schwerfällt. In dieser Arbeit zielen wir darauf ab, diese Diskrepanz sowohl aus der Perspektive des Trainings neuronaler Netze als auch ihrer Evaluierung zu verringern: Wir ermöglichen neuronalen Netzen ein tieferes Verständnis der Anatomie und zeigen, wie das gewonnene anatomische Wissen die Pathologiesegmentierung verbessert, womit sie sich dem anatomischen Schlussfolgern nach menschlichem Maßstab nähern. Abgesehen von den Fortschritten beim Training werfen wir Fragen zu den Metriken auf, die zur Messung der Performanz bei der Pathologiesegmentierung verwendet werden, und entwickeln anschließend einen neuen Ansatz, der sich stärker an der klinischen Relevanz zur Quantifizierung der Modellleistung orientiert.

Die Basis für diese Arbeit bildet ein ganzheitlicher anatomischer Datensatz, der aus einer Sammlung bereits bestehender, aber fragmentierter Datensätze zusammengestellt wird. Wir trainieren fragmentierte Modelle auf jedem der individuellen Datensätze und verwenden diese anschließend, um ihr begrenztes Wissen in einen Ganzkörper CT (Computertomografie) Datensatz in Form von Pseudolabels zu prädizieren, in dem das fragmentierte Wissen aggregiert wird. Durch iteratives Training und anatomisch orientierte Anpassungen erhalten wir einen qualitativ hochwertigen Ganzkörper-CT-Datensatz, welcher erstmals Labels für den Großteil des menschlichen Körpers enthält und auf dem das Training von neuronalen Netzen möglich ist, die weite Teile der menschlichen Anatomie segmentieren können.

Unter der Annahme, dass ein besseres Verständnis der menschlichen Anatomie es den Netzwerken ermöglicht, Pathologien und Fremdkörper besser segmentieren zu können, untersuchen wir, wie Netzwerke, die über ein anatomisches Verständnis verfügen, in zwei Aufgabenbereichen abschneiden: Segmentierung von Pathologien in ganzkörper PET/CT-Bildern und Identifizierung von Thoraxanomalien in Röntgenaufnahmen der Brust. Wir entwickeln APEx, ein neuartiges System zum gleichzeitigen Erlernen von Anatomie und Pathologie, mit Schwerpunkt auf deren wechselseitiger Beziehung. Wir verifizieren, dass ein umfassendes Verständnis der menschlichen Anatomie die Segmentierungsleistung in beiden betrachteten Aufgabenbereichen verbessert.

Neben den Fortschritten bei den Trainingsmethoden für Anatomie-Pathologie-Modelle wenden wir uns auch der Evaluierung von Modellen zu. Wir stellen fest, dass frühere Bewertungen, die weitgehend von semantischen Segmentierungsmetriken wie Dice beeinflusst sind, nur die Überlappung zwischen der Grundwahrheit und den Vorhersagen messen. Ein simpler Vergleich der globalen Überlappung greift jedoch zu kurz, wenn es um die Besonderheiten der Segmentierung von Ganzkörperläsionen geht. Da kleinere Läsionen nicht weniger wichtig sind als größere, werden diese gängigen Metriken den besonderen Herausforderungen in diesem Bereich nicht gerecht und verzerren so die Leistungsmessung des Modells. Wir adressieren diese Einschränkung durch die Entwicklung des Connected-Component (CC)-Frameworks, das beliebige Standard-Segmentierungsmetriken pro Komponente evaluiert und neu gewichtet und so die Fähigkeit der Modelle zur Segmentierung von Tumoren über den gesamten Scan hinweg unabhängig von der Tumorgröße besser misst.

Insgesamt adressiert diese Arbeit die Kluft zwischen dem numerischen und dem klinischen Verständnis der Anatomie, indem sie drei kritische technische Säulen vorantreibt: Datensatzerstellung, Modelltraining und Modellevaluation. Durch diese Beiträge wird erstens eine ganzheitliche anatomische Grundlage für neuronale Netze geschaffen, zweitens der Einfluss von anatomischem Wissen auf die Pathologieerkennung demonstriert und drittens Bewertungsmetriken entwickelt, die klinische Prioritäten besser widerspiegeln. Dadurch wird die medizinische KI in Richtung einer zuverlässigeren, anatomisch informierten und klinisch sinnvollen Zukunft vorangebracht.

ACKNOWLEDGMENTS

First, I would like to express my deepest gratitude to Rainer Stiefelhagen, my primary supervisor, for the research journey I have explored over the past four years. His kind and valuable guidance, combined with the freedom and independence to pursue my own ideas, was especially important when times were rough, particularly at the beginning of the thesis.

I am further very grateful to Jens Kleesiek for his supervision, medical expertise, and the fruitful collaborations with the University Clinicum Essen. I still have warm memories from our victory at the Eisstockschießen during our last visit to Essen. My sincere thanks also go out to Prof. Ingrisch, who has kindly agreed to review this thesis and whose work on the AutoPET dataset has significantly influenced the direction of this dissertation.

I owe a special thank you to Kailun Yang, who has been an excellent supervisor of my Master's Thesis and who sparked my interest in pursuing a research career beyond the thesis. In a second life, I will stick to the street scene segmentation field and keep working with you.

I am very grateful to the Helmholtz Imaging "HIDSS4Health" community, particularly to Nicole for her dedication to supporting her peers, and to Alexandra Walter for the enjoyable collaboration on our joint publication.

My sincere appreciation and deep gratitude go to the best colleagues one could wish for. You have made the countless hours in the lab a fun and stimulating experience. A special thank you goes to the medical team: Constantin, thanks for introducing me to the topic and for the great past and hopefully future collaborations. Zdravko, I won't forget our fun trips to Marrakesh and the impressive fireworks we watched at MICCAI. I'm always genuinely impressed by your research ideas. Simon, thanks for the many brainstorming sessions to dissect and discuss sometimes better, sometimes worse ideas. Let's keep them coming! Jiale, thanks for broadening the spectrum of our medical group; expanding our group's focus beyond segmentation is overdue. To many more successful publications!

Yufan, Junwei, and Ruiping hanging out with you guys is truly always fun. I still remember the special hot pots in Paris and all the fun nights we had, and hopefully still have. I wouldn't wanna miss the latest updates on Chinese gossip, philosophical discussions, and excellent company.

David, thanks for the constant push for bouldering and your many ideas on what to do beyond paper writing. To the successful ventures of whatever kind we think of in the future.

Furthermore, Kunyu, Jiaming, Julia, Omar, Saquib, and Angela, thank you for making the lab both engaging and sometimes challenging (Stichwort Datenschutz, Julia). Your dedication, insights, and teamwork have made coming to the lab more fun! Corinna, thanks for your patience regarding all kinds of administrative matters and your support.

Beyond academia, I would like to thank my friends from Karlsruhe: Christian, Christoph, David, Heike, and Marlon for their encouragement, guidance, open ears, and frequent fun activities, offering an often much-needed balance from work. A special thanks goes to Berlin, to Anne, thanks for many hilarious talks, fun trips, and work discussions.

A very special thanks is reserved for Betsy, whose love, friendship, perspective on life, and support have provided me with stability and support throughout the hard and stressful times.

Finally, I would like to express my heartfelt gratitude to my parents; your constant support throughout my studies is the foundation on top of which this work is built.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BIoU	Boundary IoU
CNN	Convolutional Neural Network
CT	Computed Tomography
DAP	Dense Anatomical Prediction
DICOM	Digital Imaging and Communications in Medicine
FDG	Fluordesoxyglucose
FN	False Negative
FP	False Positive
HU	Hounsfield Units
IoU	Intersection over Union
IVC	Inferior Vena Cava
JSD	Jensen–Shannon Divergence
LIMIS	Language-based Interactive Medical Image Segmentation
MAE	Mean Absolute Error
mAP	mean Average Precision
MRI	Magnetic Resonance Imaging
MSE	Mean Squared Error
NIFTY	Neuroimaging Informatics Technology Initiative
NRRD	Nearly Raw Raster Data
OAR	Organs at Risk
PEFT	Parameter-Efficient Fine-Tuning
PET	Positron Emission Tomography
PQ	Panoptic Quality
SAM	Segment Anythind Model
TP	True Positive
VLM	Vision-Language Model

Part I

BACKGROUND

1 INTRODUCTION

In this thesis, we explore how anatomical knowledge can be integrated into the three pillars of medical image analysis: dataset generation, model training, and model evaluation. While physicians develop a rich anatomical understanding through years of study and clinical experience, neural networks must derive their representations from limited and often partially labeled datasets, creating a significant gap in anatomical reasoning capabilities. To address these limitations, we first create a holistic anatomical dataset that allows neural networks to learn large parts of the anatomy rather than only a few target parts that are necessary for a specific given task. Leveraging this dataset, we address the challenging task of pathology segmentation, a task that requires physicians to rely on their understanding of human anatomy to identify unexpected structures. Finally, we turn towards model evaluation and align the prevalent usage of technically driven evaluation terms with a better clinically driven understanding in the field of lesion segmentation. Our contributions drive the field towards a more anatomically informed and clinically meaningful future.

In a famous quote, now Nobel Prize winner Geoffrey Hinton warned, "People should stop training radiologists now. It's just completely obvious within five years, deep learning is going to do better than radiologists" [110]. Today, we know that this prediction did not become reality. Radiologists are still one of the most in-demand medical specialists with a shortage of qualified personnel [1, 13, 52]. For instance, in the UK, the demand for imaging, such as MRI or CT scans, increased by 11% in 2023, the radiologist workforce, however, only increased by 6.3% despite 2023 being a particularly strong year [306]. In the US, similar trends can be observed [1], clearly indicating that radiologists continue to play a vital role in healthcare delivery.

Research demonstrates that AI models can match or exceed human radiologist performance in specific, well-defined diagnostic tasks, such as breast cancer screening and chest X-ray analysis [169, 214], indicating the potential of current technical advancements. However, these AI systems typically operate within narrow domains and cannot replicate the comprehensive clinical reasoning that radiologists bring to patient care. This raises concerns that the automated usage of AI systems may lead to systemic errors with high consequences [82, 109, 356] and highlights current trends that point towards the direction of a fruitful co-existence [17, 44, 169, 246] of radiologists with AI models serving as specialized tools supporting the decision-making.

Contrary to the narrow specialization of current AI systems, human radiologists integrate imaging findings with patient history, clinical context, and have detailed anatomical reasoning regarding the entire human anatomy to make nuanced diagnostic decisions that extend beyond task-specific objectives and limited anatomical knowledge.

Within this thesis, we draw inspiration from the comprehensive anatomical reasoning that radiologists possess and aim to equip neural networks with a broader anatomical knowledge to enhance their diagnostic capabilities through more comprehensive anatomical understanding, moving beyond the narrow, task-specific focus that currently limits their clinical utility. We pursue this goal through a systematic approach by following the model development lifecycle that addresses three fundamental aspects of medical AI: first, we create a holistic anatomical dataset that enables networks to learn large parts of human anatomy rather than isolated structures; second, we develop training methodologies that leverage this anatomical knowledge to improve pathology detection; and third, we establish model evaluation frameworks that better align with clinical priorities and meaningful assessment of model performance.

1.1 THE ROLE OF SEGMENTATION MODELS IN CLINICAL SETTINGS

Within this thesis, we aim to tackle the task of semantic segmentation for both anatomical and pathological structures. Segmentation models produce pixel- or voxel-wise label maps, assigning each location of the input image to a predefined anatomical or pathological class. This adds spatial semantic context to the image, supporting accurate diagnosis, interpretation, treatment planning, and quantitative disease monitoring [204]. The usage of segmentation models in the clinical field is very versatile.

In *oncology*, models could be used for direct segmentation of various types of cancers [231, 277, 359] or lung nodules [6], detecting lesions and supporting oncologists in the diagnostic workflow by providing features such as shape, morphologic structures, or texture [111]. Beyond the direct segmentation of cancer, an important task in the treatment of patients with radiation therapy is the segmentation of organs-at-risk. These are potentially radiation-sensitive organs that are close to the treatment area, whose accurate delineation is required to customize the radiation therapy to the patient's anatomy. The usage of automatic segmentation approaches [314] has shown great potential for improving accuracy, efficiency, and consistency in image-guided radiotherapy [200, 267]. After treatment, automatic segmentation models can support longitudinal disease monitoring by quantifying tumor volumes over time [62].

Besides oncology, automated segmentation approaches have been successfully tested across various other clinical disciplines. In *Neurology*, automated segmentation approaches could identify multiple sclerosis lesions [162] or support in stroke detection by segmenting intracranial hemorrhage [172]. In *Cardiovascular Imaging*, segmentation models were successfully applied to echocardiogram videos, which can support preliminary interpretation in areas with insufficiently qualified cardiologists [83]. The field of *Thoracic Imaging* has seen a rapidly growing interest in AI-based tools, particularly during the COVID-19 pandemic [243]. Segmentation models can be used to either directly segment abnormalities [40, 69] or used as preprocessing tools, e.g., by preselecting the area of interest in CT scans before applying detection models [176].

Automated segmentation tools also enable the collection of biomarkers, which can be used for downstream analysis such as pancreatic CT attenuation and visceral fat for type 2 diabetes mellitus [302], automatic bowel measurements for biomarker extraction related to constipation [245] or whole-body MRI segmentation of visceral fat, subcutaneous fat, and muscle mass to assess a patient's metabolic health and nutritional status, which are indicators for cardiac diseases, type 2 diabetes mellitus, and cancer [166].

In summary, the use of segmentation models has the potential to directly impact clinical workflows across multiple disciplines by supporting physicians in a wide range of tasks.

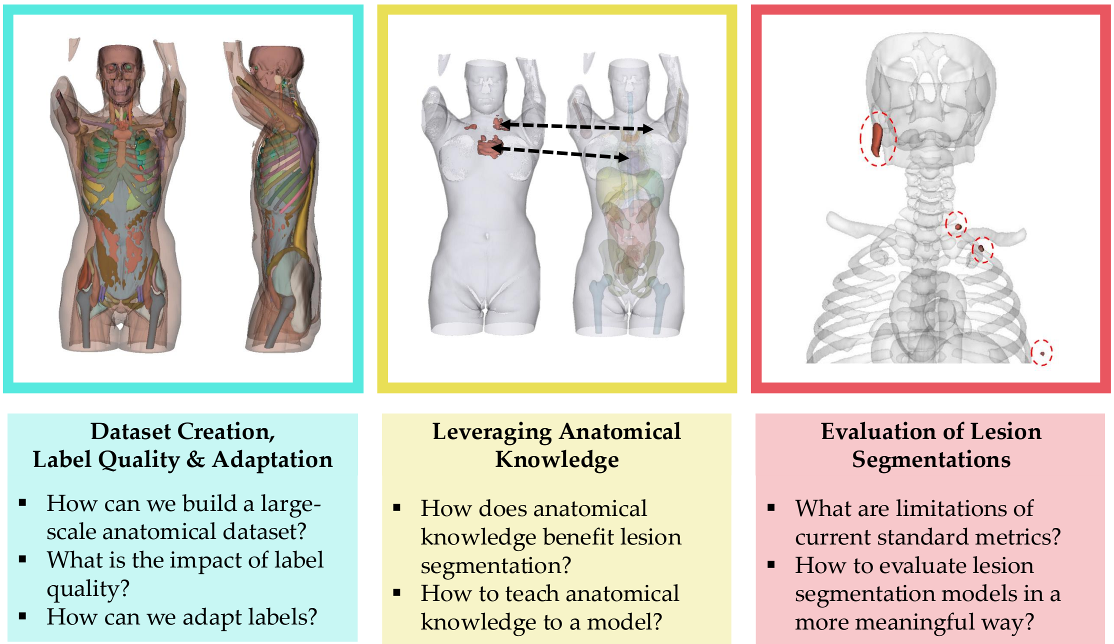
1.2 THESIS ROADMAP AND CONTRIBUTIONS

In this dissertation, we seek to integrate anatomical knowledge into the development lifecycle of medical imaging models. We discuss related work in Chapter 2 and place our contributions within the field. In Chapter 3, we explore the process of automatically generating a labeled anatomy dataset and examine the implications of using automated labels. Chapter 4 discusses how anatomical knowledge in the form of labels can aid the segmentation of pathologies. To complete the model life-cycle, we discuss how lesion segmentation models should be evaluated in Chapter 5. We conclude this thesis by summarizing our insights in Chapter 6 and pointing to future research directions in Chapter 7. A complete list of publications that resulted from the work on this thesis can be found in Appendix B.

In the following, we briefly summarize our contributions across Chapter 3, Chapter 4 and Chapter 5.

1.2.1 *How to Create Anatomical Labels without Medical Supervision?*

Supervised learning remains the most popular learning scheme for training biomedical neural networks. It is a task that requires pixel or voxel-wise annotations. To enable a holistic understanding of the human anatomy, a network would require a large number of annotated images as well as a large number of annotated anatomical structures per image. Tasking doctors to annotate scans from scratch is infeasible due to the



Dataset Creation, Label Quality & Adaptation

- How can we build a large-scale anatomical dataset?
- What is the impact of label quality?
- How can we adapt labels?

Leveraging Anatomical Knowledge

- How does anatomical knowledge benefit lesion segmentation?
- How to teach anatomical knowledge to a model?

Evaluation of Lesion Segmentations

- What are limitations of current standard metrics?
- How to evaluate lesion segmentation models in a more meaningful way?

Figure 1: Overview of the proposed contributions across the model development lifecycle: From left to right: 1) **Dataset Generation**: We generate a holistic anatomical dataset automatically, assess the impacts of label quality, and investigate dynamic label adaptation. 2) **Pathology Model Training**: We investigate strategies to incorporate anatomical knowledge to enhance the segmentation of pathologies. 3) **Model evaluation**: We address how lesion segmentation models can be evaluated with better metrics.

massive time required and the high expected costs.[1] We describe our approach to address this challenge in Chapter 3. We leverage existing datasets with a low number of annotations and aggregate their anatomical information via neural-network-based pseudolabeling into an empty full-body PET/CT dataset as presented in our *ICIP 2024* publication [128].

A limitation of this approach is that, although our dataset was generally rated as impressive by radiologists, we are unable to evaluate the pixel-wise accuracy of each individual mask. Thus, when training segmentation models on the dataset, this limitation may result in training on imprecise masks. As this problem is not confined to our dataset, but a more common issue in large datasets [20, 177], we analyze the influence of label quality on various tasks based on our work [127] *(currently under submission).*

Modern foundation models [122, 130, 204, 334, 337], trained on large-scale collections of datasets, inevitably inherit labeling errors. Moreover, even with perfect annotations, segmentation models can produce inaccuracies in clinical settings. To address such imperfections, we investigate how physicians can interact with models that were potentially trained on noisy labels, and iteratively refine low-quality predictions using natural language commands. This approach is embodied in the LIMIS architecture, presented in our *ISBI 2025* publication [105].

1.2.2 *How to Leverage Anatomical Labels for Improved Pathology Segmentation?*

Besides the segmentation of anatomical structures, applications in oncology demand the accurate delineation of tumor volumes. Given the intrinsic interconnection between anatomical and pathological tissue, we hypothesize that anatomical knowledge may have beneficial effects for segmenting pathologies. In Chapter 4, Section 4.1, we build upon this hypothesis and explore various strategies to include anatomical knowledge into the training process of a Mask2Former [48] based architecture to segment pathologies in PET/CT imaging and thoracic abnormalities in chest X-ray. We develop a dual-decoder architecture that forces the model to segment anatomies and pathologies jointly. We interleave the two decoders with the possibility to exchange information in our final Anatomy-Pathology Exchange (APEx) architecture, which leads to an improved performance for pathology segmentation across both domains (published in *MICCAI 2024* [128]).

APEx, however, requires a modified architecture and remains limited to a 2D formulation due to its focus on the X-ray and PET/CT domain, which constrains its direct applicability in volumetric workflows. In the subsequent Section 4.2, we investigate whether comparable anatomy-guided improvements can be achieved without additional supervision or re-engineering of the target model. This leads to the GRASP framework, published at the MLMI workshop at MICCAI 2025 [180], which directly

[1] Assuming an optimistic 10 minutes per mask and 60€ per hour for a radiologist, our dataset would cost more than 600k €

injects anatomical priors from frozen, pre-trained anatomy models into the pathology training process via pseudo-labels and bottleneck-level feature fusion. GRASP eliminates the need for auxiliary anatomical losses, operates natively in 3D, and leverages existing anatomy segmentation networks as fixed knowledge sources, enabling anatomy-aware pathology segmentation without relying on anatomical learning as a proxy loss.

1.2.3 *How to Evaluate Pathology Segmentation Models in a Better Way?*

Although the segmentation of pathologies such as lesions is commonly approached as a semantic segmentation problem by the research community, identifying individual metastases remains highly relevant. From a medical perspective, it can make a significant difference whether the same volume of cancerous tissue is confined to a single metastasis or spread between multiple metastases. Yet, from a semantic segmentation evaluation perspective, the most important metric is measuring overlap, thereby ignoring topological structures. In Chapter 5, we discuss limitations of evaluating the performance of lesion segmentation models using standard semantic segmentation metrics, as well as established instance-aware semantic segmentation metrics such as Panoptic Quality [153]. We develop the CC-Metrics evaluation protocol, which addresses the identified limitations and allows the usage of well-established standard segmentation metrics on a per-connected component basis (published in *AAAI 2025* [129])

A Note on Implementation:
Alexander Jaus is responsible for the implementation of all frameworks developed in Section 3.1, Section 3.2, Section 4.1 as well as Chapter 5. Two sections are based on joint works resulting from very close collaborations with his Master's Thesis students: Lena Heinemann (Section 3.3) and Keyi Li (Section 4.2). Both Alexander Jaus and the corresponding student have made substantial contributions to the research in the mentioned section. While it is difficult to set a precise boundary, Alexander Jaus was rather in charge of the idea, while the student focused on the implementation. Section 3.1 stems from a collaboration in which Alexander Jaus was, as previously mentioned, rather responsible for the implementation, whereas Constantin Seibold contributed more to the idea.

2 RELATED WORK

The developments in this thesis broadly fall at the intersection of computer vision and radiology, with a particular focus on the image domain of PET/CT and the technical task of segmentation. Our contributions advance several topics and can be roughly clustered into three distinct areas: Data-centric contributions, such as the development of anatomical datasets and model training under noisy labels; Model-centric innovations in the field of interactive segmentation and the inclusion of anatomical priors for pathology segmentation; and finally, we contribute to the field of evaluation metrics. The following chapter surveys the relevant literature in each of these areas and contextualizes our contributions within the broader research landscape.

2.1 ANNOTATED ANATOMICAL DATASETS

Any supervised training approach relies on a dataset containing annotated examples of anatomical structures from which a learner, such as a neural network, can derive mappings from images to expected labels. In a fully annotated dataset, for each image exists at least one mask containing pixel- or, in the case of volumetric images, voxel-wise annotations for one or multiple target structures of interest. While the technical details of how this mask is stored vary and are not of primary interest in this work, the masks establish a spatial relationship between the image and a semantic understanding of the image.

In the following, we explore the development of datasets in the domain of CT, as this is of primary interest in this work. Adjacent areas such as MRI or X-ray have seen comparable progress along with the CT domain. Whenever feasible, we reference the publication associated with the release of the datasets; if this is not feasible, we directly cite the data source.

2.1.1 *Single Organ Dataset*

Initial datasets primarily focused on segmenting individual organs in CT scans, often using relatively small sample sizes. The development of liver segmentation algorithms, for instance, began with datasets comprising only 20–30 expert-annotated cases [104, 295, 310] and increased later to larger databases that also included tumors and data from multiple medical sites [28]. Other datasets with single organs of interest include the pancreas [269], the kidneys [106], the spleen [14, 15, 41], pulmonary-vessels [274], the lungs [196, 228], airway-paths [195, 367] and the hearth [383]. The latter three datasets, however, already include multiple sub-structures such as individual lung lobes or more specific labels within the airways and the heart, which already adds more semantic context for a specific organ of interest. Generally, with the increasing performance of segmentation algorithms, datasets have become more complex, and multi-organ datasets have gained popularity, encompassing multiple anatomical structures.

2.1.2 *Multi Organ Dataset*

While the analysis of single organs may be of interest for specific diseases, such as liver- [28] or kidney tumors [106], the applications of models trained on such datasets are confined to the specific organ. While efforts such as the Medical Segmentation Decathlon [5, 294] aim to expand the generalizability of networks across a number of distinct tasks and modalities by training and testing networks on multiple distinct tasks, such as liver-tumor segmentation or colon segmentation, each task remains confined to a specific organ or disease of interest.

Multi-organ datasets offer an intuitive approach to enhancing the capability of segmentation models. Each scan in a multi-organ dataset contains annotations for several anatomical structures. Early approaches to multi-organ datasets include the works of Lee et al. [171], which provides segmentations for 12 structures in 20 CTs to evaluate registration-based segmentation approaches. This work is later extended to a dataset of 100 CTs [349] and 13 structures. The BTCV dataset [168] is a notable example, containing 30 abdominal CT scans with annotations for more than 10 abdominal structures, including the liver, spleen, kidneys, glands, pancreas, gallbladder, esophagus, stomach, and vessels. Due to its broad coverage of the abdomen, it remains a popular dataset today, despite its limited size.

Chaos [148] is an abdominal dataset including MRI and CT data on healthy individuals with liver, kidney, and spleen annotations. AMOS [133] provides 500 CTs with 15 annotated abdominal structures, while FLARE [205] combines 2000 unlabeled with 50 labeled cases to develop semi-supervised approaches for 13 abdominal organs.

To increase dataset sizes and diversity, researchers have started to build upon previously published datasets by extending one or multiple datasets with additional annotations. CT-ORG [265] builds upon the LITS [28] dataset containing liver-related annotations and extends it to 6 organs: liver, lungs, bladder, kidney, bones, and brain. CT1K [206] builds upon multiple existing datasets [28, 106, 269, 294] containing single organ annotations to build a large abdominal dataset of 1112 CT scans. They find that while deep-learning approaches already perform well for normal cases, unseen or rare diseases remain challenging. Their final dataset contains annotations for liver, kidney, spleen, and pancreas.

Gibson et al. [85] combine the BTCV [168] and pancreas datasets [269] and extend it to a common labeling scheme containing spleen, kidney, gallbladder, esophagus, liver, stomach, pancreas, and duodenum.

A key challenge that arises with the desire to create larger and larger datasets is the increased workload for annotators. The required expertise to qualify for annotating medical images further complicates scaling annotations to unskilled workers, e.g., via Amazon Mechanical Turk (AMT), which has been widely used in the natural image domain [58, 161, 183, 343]. One approach that has been on the rise, besides building on top of existing labels, is to generate human-in-the-loop systems that aim to minimize human involvement by leveraging pre-trained segmentation models, generating segmentation that experts only need to correct. Qu et al. [256] present Abdomenatlas8k, a dataset consisting of over 8k CTs with annotation masks for spleen, liver, kidneys, stomach, gallbladder, pancreas, aorta, and IVC by pointing experts to uncertain regions of automated annotations. Using a similar technique, Abdomenatlas 1.1 provides 25 anatomical structures on over 9k CT scans [179], which is further extended by the same author [177].

Outside the abdomen, the Auto-segmentation [259] challenge dataset aims to benchmark models for segmenting head-neck structures, which can be beneficial for radiation therapy. The dataset consists of 40 images covering the brainstem, mandible, chiasm, bilateral optic nerves, bilateral parotid glands, and bilateral submandibular glands. Walter et al. [315] provide a head-neck dataset targeted to aid radiotherapy with 71 annotated structures on 104 CT scans. SegTHOR [167] focuses on 4 organs in the thoracic region, CTPelvic1K [187] provides annotations for lumbar spine, sacrum, left hip, and right hip on over 1000 CTs, marking an important step towards pelvis segmentation, while PENGWIN [192, 193] focuses on the segmentation of pelvic fractures. Verse [282] provides accurate masks for individual vertebrae, and SpineMets [250] is another verse dataset providing next to vertebrae, spine-related tumor masks. RIBSEG [137, 354] segments individual ribs, and Pediatric [139] addresses the segmentation of 29 structures in 359 pediatric chest-abdomen-pelvis and abdomen-pelvis CTs.

Besides the addition of more labels on a larger number of scans, other approaches develop datasets that expand beyond standard body regions, such as the abdomen, pelvis, or thorax, offering annotations on full-body scans.

The VISCERAL anatomy benchmark [136] allows model evaluation on 30 whole-body contrast and non-contrast CT and 30 MRI scans for 20 structures in 15 organs. TotalSegmentator [333] annotates 104 anatomical structures spanning the entire body. SAROS [157] follows a different approach, providing annotations for distinct body regions and tissues such as the thoracic cavity, bones, brain, breast implant, or mediastinum, thereby focusing on less common labels.

An overview of the most relevant datasets which have been discussed in this section is shown in Table 1, including our own contribution, the DAP-Atlas Dataset [128].

Table 1: Overview of anatomy CT datasets including their focus, number of CT scans, number of annotated structures, and publication year

Dataset	Body Region	# CTs	# Structures	Year
Partial Body Datasets				
Registr. Challenge [171]	Abdomen	20	12	2015
Registr. Challenge [349]	Abdomen	100	13	2016
BTCV [168]	Abdomen	30	13	2015
Chaos [148]	Abdomen	80	4	2021
AMOS [133]	Abdomen	500	15	2022
FLARE [205]	Abdomen	2050	13	2023
CT1K [206]	Abdomen	1112	4	2021
V–Network [85]	Abdomen	90	8	2018
AbdomenAtlas-8k [256]	Abdomen	8000+	8	2023
AbdomenAtlas 1.1 [179]	Abdomen	9000+	25	2024
Auto-seg. [259]	Head/Neck	40	9	2017
Head–Neck OAR [315]	Head–Neck	104	71	2024
SegTHOR [167]	Thorax	60	4	2020
RIBSEG [354]	Thorax	660	24	2021
CTPelvic1K [187]	Pelvis	1000+	4	2021
PENGWIN [192, 193]	Pelvis	150	4	2025
Verse [282]	Spine	374	28	2021
SpineMets [250]	Spine	55	24	2024
Pediatric [139]	Chest–Abd.–Pelvis	359	29	2022
CT–ORG [265]	Mixed	140	6	2020
Whole Body Datasets				
VISCERAL [136]	Whole body	30	20	2016
TotalSegmentator [333]	Whole body	1204	104	2023
SAROS [157]	Whole body	900	19	2024
Our Contribution				
DAP Atlas [130]	Whole body	533	142	2023

We further discuss the trade-off between dataset size and annotation comprehensiveness in Figure 2. This scatter plot shows the relationship between dataset size (number of CT scans) and annotation comprehensiveness (number of anatomical structures) across existing datasets in the literature. Both axes use logarithmic scales to accommodate the wide range of values. Datasets are categorized into three groups: partial body datasets (blue circles) focusing on specific anatomical regions such as abdomen, thorax, head/neck, pelvis, or spine; whole body datasets (green circles) providing comprehensive anatomical coverage; and our contribution, DAP Atlas (orange circle), which achieves the highest number of annotated structures (142) while maintaining a substantial sample size (533 CT scans). The plot reveals that most existing datasets face a trade-off between sample size and annotation comprehensiveness, with larger datasets typically annotating fewer structures due to the increased labeling effort required.

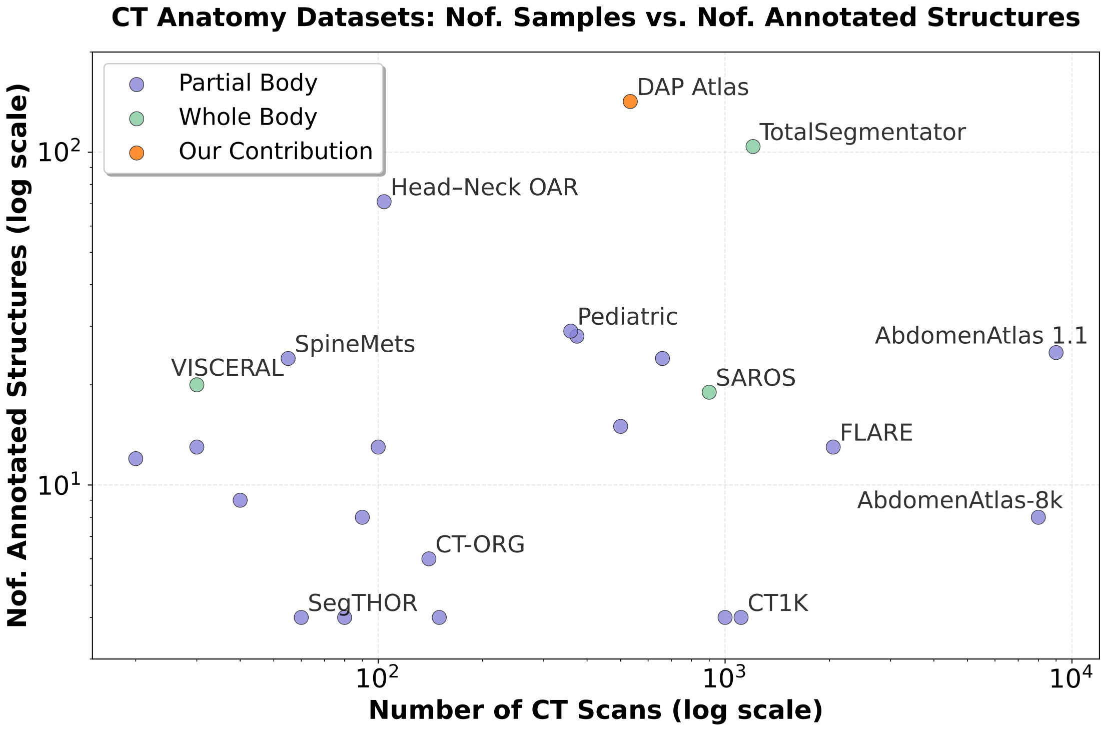


Figure 2: Comparison of CT anatomy datasets by number of samples versus number of annotated structures (both axes in log scale). Partial body datasets (blue) focus on specific anatomical regions, whole body datasets (green) provide comprehensive coverage, and our DAP Atlas (orange) achieves the highest number of annotated structures (142) with substantial sample size (533 CT scans) on whole-body scans.

Our contribution:
Analysis of existing datasets reveals three fundamental limitations in the current dataset landscape: (1) Limited dataset size: Early datasets typically contain fewer than 100 annotated CT scans, constraining model training capabilities. (2) Limited anatomical coverage: Most datasets focus on specific body regions such as the abdomen or pelvis, preventing comprehensive whole-body segmentation. (3) Limited structural annotation: The number of annotated structures per dataset remains low due to the substantial manual effort required for volumetric annotation. These limitations collectively hinder the development of segmentation models capable of comprehensive human body analysis. To address these challenges, we propose the DAP-Atlas dataset in Section 3.1, which provides the most extensive structural annotations (142 structures), covers the entire body, and maintains substantial scale (533 CT scans).

2.2 TRAINING SEMANTIC SEGMENTATION MODELS ON IMPERFECT LABELS

The desire to create larger and larger datasets poses a significant challenge: the increased workload for annotators. Moreover, the required expertise to annotate medical images further complicates the scaling of annotations. In high-stakes clinical settings such as OAR segmentation for radiation therapy, the deviation of structures must follow strict guidelines such as the ones for head-neck radiotherapy [91, 92]. As segmenting anatomical structures by hand leads to intra- and inter-rater variability [33, 211, 238], fusion approaches such as STAPLE [332] have become the gold standard. Segmenting by hand, let alone having multiple radiologists whose segmentations can be fused, becomes increasingly less feasible as dataset sizes grow. This has led to the development of automated [130] and semi-automated dataset creation approaches [133, 177, 179, 205, 256, 265, 333] that utilize human-in-the-loop systems, aiming to minimize human involvement by leveraging pre-trained segmentation models whose predictions are corrected by experts. The focus of segmentation quality assessment shifts from merely annotating labels to identifying errors. Wasserthal et al. [333] address this challenge through manual review processes, in which physicians meticulously examine the segmentations and implement corrections upon detecting errors. Other approaches utilize model uncertainty [177, 256] to direct reviewers to regions of high uncertainty. This approach relies on the assumption that errors can be explained by the model's uncertainty measures. While the obtained annotation speedup is impressive and the field will likely continue moving in that direction, a key question remains difficult to answer: How good are these labels actually? First works [177] point out that there are larger-than-expected errors in the popular TotalSegmentator [333] dataset, particularly for difficult classes such as the colon. The same is true for the AbdomenAtlas dataset [20].

This prompts critical research questions:

- What is the influence of annotation errors on the efficacy of model training?
- In what ways does label quality impact downstream performance across various tasks?

These questions are highly relevant in the current landscape of data-centric medical AI. Within this section, we review works that have examined the impact of label quality on model performance. This is most commonly explored in robustness studies, where authors propose and evaluate robust learning schemes. These will be briefly reviewed in the following. The focus of this section is on medical segmentation tasks; however, if deemed appropriate, we will also relate to adjacent tasks, such as classification, and papers in the natural image domain.

As we focus on the task of supervised learning, we examine datasets containing dense masks for each image. Semi-supervised learning [379, 380] approaches although sharing certain similarities with label-noise approaches, since some techniques rely on pseudo-labeling unlabeled images for self-training [18, 69, 95, 203, 254, 307, 324, 340, 346, 373], or co-training [77, 164, 218, 247, 254, 317, 322, 342, 376] which inherently introduces label noise are not part of this review. We refer the interested reader to a comprehensive overview by Han et al. [96]. The key distinction between semi-supervised setups and the research questions we address lies in the different fundamental training paradigm: while semi-supervised learning assumes access to a clean labeled subset alongside unlabeled data, our focus examines scenarios where the entire labeled dataset may contain annotation errors from the outset. More critically, we are concerned with measuring the expected performance implications when researchers adopt these publicly available datasets off-the-shelf to train or pre-train models on potentially noisy labels, without prior knowledge of the underlying annotation quality.

Other forms of imperfect data, such as learning with weak labels [119, 325], and images with measurement noise [147, 234, 260, 275, 365], have been addressed in the literature but are beyond the scope of this work. Within the following, we build upon the taxonomy established by Shi et al. [286].

2.2.1 *Robust Network Learning Approaches*

Inspired by the progress in the field of training networks under noisy labels in the natural domain [24, 170, 224], Dgani, Greenspan, and Goldberger [60] investigate the performance of models under label noise in the task of microcalcification classification. They add a layer at the end of their CNN which aims to aid in the identification of the unobservable true label by modeling a noise transition function. In a similar fashion, other works model confusion matrices, estimating label confusions for each annotator [278, 298, 305, 366]. Given that labels are typically hard labels, in the sense that they assign 100% certainty to a single class, several approaches have introduced label smoothing as a form of regularization. Label smoothing [123, 227, 248] replaces the one-hot target with a softened version that assigns a small probability to incorrect classes, reducing overconfidence and improving generalization for X-ray classification [249], real-time endoscopy segmentation [252], and fetal brain segmentation in ultrasound imaging using an approach that reflects spatial and anatomical uncertainty [145]. In addition to mitigating overconfidence in labels through label smoothing, modifying the training loss function represents another effective strategy for addressing noisy labels. Mean Absolute Error (MAE) [84] and its improved variant, iMAE [327], were originally introduced in the context of natural image processing. These methods reduce the influence of potentially incorrect labels compared to conventional cross-entropy loss functions or Mean Squared Error (MSE). Karimi et al. [144] explore these techniques in the domain of brain lesion segmentation in MRI imaging, histopathology classification, and fetal brain segmentation. Wang et al. [316] adapt the commonly used Dice Loss [219], which is an MSE-type loss by reformulating it into an MAE-type loss, showing increased robustness for pneumonia segmentation in CT images. Other works derive robust losses by maximizing the log-likelihood of a Student t-distribution [89] and validate on skin-lesion segmentation and lung-segmentation in X-ray images.

In addition to developing novel robust loss functions, researchers have proposed leveraging conventional loss functions while modelling the contribution of samples suspected to be noisy by assigning them reduced weights during training. Determining the likelihood of labels being noise can be done based on the probability of the data being an outlier [351] using algorithms like Local Outlier Factor (LOF) [160], learning-based reweighing schemes [378], Gaussian-filter based [344], meta-learning based [222] and uncertainty based [140].

Excluding data that appears to be noisy directly is a further strategy. This can be achieved by examining loss magnitudes [331], building on top of the observation that noisy samples tend to yield higher losses relative to clean samples [9]. Other strategies involve leveraging techniques from semi-supervised learning, such as Co-Training [12, 70, 220, 350], which highlights inconsistencies among different networks.

Besides presenting innovative methods for addressing robustness and conducting noise performance analysis, a limited number of studies have sought to quantify the impact of label noise for standard model training. Yu et al. [360] investigate the impact of label noise in CT scans on mandible label annotations performed by multiple experts. The annotations are compared to enhance label quality, although the training process utilizes the original, potentially flawed labels. Brückner et al. [34] conducted a small analysis of two types of noise: random boundary noise and systemic over- or under-segmentation, in the context of abdominal organ segmentation on CT imaging.

We summarize related work on label noise in Table 2, focusing on CT segmentation methods.

	Task	Domain	# Datasets	Medical dataset(s)
		2D-based studies		
Zhang [366]	S	CT, MRI	1,2	MSLesion [297] BraTS [216] LIDC-IDRI [8]
Schmidt [278]	S	H	3	Gleason [237] Arvaniti [11] CrowdSeg [3]
Pornvoraphat [252]	S	E	1	private
Wang [316]	S	CT	1	private
Gonzalez [89]	S	DER, X-ray	1,1	ISIC [56] Shenzhen [126]
Zhu [378]	S	X-ray	1	JSRT [289]
Xiao [344]	S	H	1	Gleason [237]
Mirikharaji [222]	S	DER	1	ISIC [56]
Wang [331]	S	CT	1	Spine-CT [358]
Jin [140]	S	CT	1	SegThor [167]
		3D-based studies		
Sudre [298]	C, D	MRI	1	private, based on [308]
Liao [182]	D	CT	2	LIDC-IDRI [8] LungX [7]
Karimi [145]	S	MRI	1	private
Karimi [144]	S, C	MRI, H	2,1	2 × private Gleason [237]
Yu [360]	S	CT	3	private 2 × Head–Neck [236, 259]
Bruckner [34]	S	CT	1	CT-ORG [265]
Fang [70]	S	CT, MRI	1,2	LIDC-IDRI [8] ACDC [27] BraTS [216]
Ours	S	CT	5	BTCV [168] WORD [202] AMOS [133] CT1K [206] AbdomenAtlas [177]

Table 2: Overview of most relevant related work on noisy-label handling in medical imaging. Task: Classification (C), Detection (D), Segmentation (S). Domain: CT, MRI, Histopathology (H), Endoscopy (E), Dermatology (DER). #Datasets denote the number of datasets per medical domain as denoted in the previous column.

Despite extensive research on label noise, volumetric CT segmentation remains underexplored. While the majority of works focus on 2D approaches in the first place, thereby ignoring inter-slice relations, existing volumetric approaches suffer from several limitations: (1) evaluation on single or very few CT datasets [34, 70], (2) focus on narrow anatomical structures like mandibles [34], (3) reliance on artificially generated noise that may not reflect real-world annotation errors [34, 70, 89, 331, 378], and (4) dependence on multiple annotators, which is impractical for large-scale studies [144, 145, 360].

Many proposed methods introduce complex domain-specific knowledge [145], optimization procedures [278], or architectural components [316] that hinder general applicability.

Our contribution:
We address the identified gaps by providing the first comprehensive analysis of label quality effects in large-scale CT segmentation datasets, thereby focusing on the realistic case where researchers use datasets with unknown label noise, treating them as if they were clean labels. (1) We examine different levels of noise in large dataset generation processes via multi-foundation model pseudolabeling, showcasing realistic noise patterns present in novel semi-automatically generated datasets. (2) We analyze the trade-off between dataset size and annotation quality, thereby providing guidance for both dataset providers and users. (3) We conduct, to our knowledge, the first systematic study of label quality effects for the scenario of pretraining models on noisy volumetric CT segmentation. (4) We offer practical recommendations for leveraging large, potentially noisy datasets in medical imaging applications.

2.3 INTERACTIVE SEGMENTATION APPROACHES

A key limitation of standard supervised models is their lack of controllability: they act as deterministic functions, producing fixed outputs for a given set of inputs. This becomes problematic when training data contains imperfect labels, as the model may internalize incorrect information. Of course, models can also make errors even when trained under ideal conditions, due to factors such as limited capacity, overfitting, or inherent ambiguity in the data. Moreover, even under the assumption of perfect labels and perfect prediction capabilities, discrepancies between the training and inference annotation protocols may lead to misalignment in model behavior [185, 309]. Interactive models offer a promising alternative by enabling dynamic control over the prediction process. Rather than producing fixed outputs, these models can adapt their behavior in response to user input or contextual cues. This flexibility is particularly valuable in settings with uncertain or noisy supervision, as it allows the user to guide or correct the model's predictions. Furthermore, interactive systems can bridge mismatches between training and deployment protocols by incorporating real-time feedback, thereby increasing robustness and practical utility.

Interactive segmentation methods have roots in the pre-deep learning era, with early approaches like interactive graph cuts and GrabCut establishing foundational techniques for user-guided image segmentation [31, 72, 270]. With deep learning models dominating segmentation in recent years, initial attempts combined interactive segmentation with deep learning models [208, 347]. However, only with the introduction of the Segment Anything Model (SAM), the field of interactive segmentation experienced a significant increase in interest, as the SAM model combined large-scale dataset training with promptable segmentation of arbitrary objects [154].

In the medical domain, a similar trend is evident. Early deep learning–based methods demonstrated the utility of interactive models for clinical applications [4, 32, 173, 257, 276, 374], and more recently, the field has also benefited from the broader surge of interest in interactive segmentation [210]. While the original SAM model was shown to have a decent performance in the medical field as a zero-shot segmentation model [213, 273], SAM-based adaptations were quickly developed, ranging from Parameter-Efficient Fine-Tuning (PEFT) adaptations [50, 88, 285, 318, 341] to full adaptations [204].

A key design feature of interactive models is the type of interactions the models can handle: Typically, these are physical interactions such as clicks [181, 188, 276, 299], scribbles [4, 16, 51, 382], bounding boxes [74, 257, 304] or even multiple of these interaction types [45, 159, 184, 337]. A key type of interaction, however, is missing: natural language. Text-based guidance for vision models has picked up pace for open-set object detectors with pioneering works such as OV-DETR [363], GLIP [76], or Grounding DINO [191] and segmentation networks with LAVT [357], CRIS [328], X-Decoder [384], or Grounded-SAM [264]. In the medical field, text-guided segmentation has been tested for surgical instruments [377], endoscopy [29], and radiological imaging [158, 185, 258]. However, these approaches face a critical limitation: while they enable text-based segmentation, they lack the capability to perform extended interactions beyond initial segmentation using natural language, forcing users to rely on physical interactions (clicks, scribbles) for refinement: an impractical requirement in clinical scenarios where physicians' hands must remain free for surgical procedures, patient care, or equipment operation.

Our contribution:
We introduce LIMIS: The first purely language-based interactive medical image segmentation framework, addressing the limitation that existing methods require physical interactions, which are impractical in clinical scenarios where physicians' hands are occupied. Our key contributions are: (1) A Medical Language-to-Segmentation Pipeline: We adapt Grounded SAM to medical CT images by fine-tuning Grounding DINO via LoRA, enabling initial mask generation from

natural language prompts. (2) Language-Only Interaction Loop: We pioneer a completely hands-free segmentation refinement loop through natural language commands, supporting both manual adaptations and automated multi-step strategies for common medical segmentation errors. (3) Clinical Validation: We demonstrate the effectiveness of our approach across three medical datasets and validate the system's usability with professional radiologists.

Our work shifts interactive medical segmentation from requiring physical input to enabling hands-free, language-driven refinement, opening new possibilities for intraoperative and real-time clinical applications.

2.4 ANATOMICAL PRIORS FOR PATHOLOGY SEGMENTATION

Segmenting anatomical structures in medical images presents a distinct set of challenges compared to segmentations in natural images, including ambiguous boundaries, low contrast, limited annotated data due to high labeling costs, and severe class imbalance. Additionally, image modalities such as CT or MRI have domain-specific artifacts and noise characteristics that complicate segmentation tasks.

Medical image segmentation, however, benefits from a unique and powerful advantage: the well-established context of human anatomy. Unlike other imaging domains, medical images capture structures that, to some extent, follow predictable anatomical patterns and relationships, providing a rich foundation of prior knowledge that can be systematically leveraged to improve segmentation accuracy. This anatomical consistency makes medical imaging particularly well-suited for incorporating domain-specific knowledge into segmentation algorithms. In the following section, we explore various strategies for integrating anatomical priors into the segmentation of normal anatomical structures, before examining how this anatomical understanding can enhance the detection and delineation of pathological tissues.

2.4.1 *Anatomical Knowledge for Anatomy Segmentation*

Prior to deep learning, models such as active contour models [146], level-set methods [59], or even simple threshold methods, like Otsu's method [242], aimed to directly model decision boundaries. These approaches allowed for a direct incorporation of anatomical knowledge through the model's parameters [93, 244, 287], geometric and topological constraints via (multi) atlas-based segmentation [19, 385] or knowledge about the expected tissue appearance [290]. However, with the rise of deep learning approaches, particularly pixel-wise optimization methods such as fully convolutional networks [197], such as U-Net architectures [268], the incorporation of anatomical knowledge has become less straightforward. The deep learning paradigm's promise to learn complex patterns directly from data, combined with its demonstrated superior performance in many applications [120], raises questions about the continued relevance and merit of explicitly incorporating anatomical information.

Some approaches decide to combine the power of deep-learning approaches with the flexibility of classic learning regimes to model the incorporation of prior anatomical knowledge, e.g., via initial segmentations using deep learning and refinement via anatomy-informed classic segmentation methods [116, 291, 362, 371] or direct incorporation of classical methods within the deep learning architecture [261, 303, 372].

A more straightforward approach to incorporating knowledge into deep learning models is to adapt or extend the loss functions. Traditional segmentation losses like Cross-Entropy and Dice Losses [219] are inherently limited to data-driven optimization, treating voxels independently or maximizing spatial overlap. Focusing solely on data optimization fundamentally limits the incorporation of prior anatomical knowledge or geometric constraints. This limitation is typically addressed through the extension of the loss function by a regularization term into which prior objectives, including shape [79, 221, 239, 361], topology [55, 152], size [150], or interrelations [25] are embedded.

In addition to regularizing models, learning meaningful properties can be achieved through multi-task prediction. This approach explicitly directs the model's attention to beneficial properties, such as edges [42] or geometric features, for instance, via distance transforms [233, 326] or atlas-based image registration [132, 348].

A further direction is leveraging the model architecture directly to model knowledge using Graph Neural Networks [134, 283, 284, 288, 329, 370], generative models [57], or Bayesian learning frameworks [117].

2.4.2 *Anatomical Knowledge for Pathology Segmentation*

In contrast to the extensive incorporation of anatomical knowledge for the segmentation of anatomical structures, the application of such knowledge for pathology segmentation remains less explored. A few works have leveraged anomaly detection through reconstruction approaches such as diffusion models [26, 336], VQ-VAE [251], or regression-based approaches [35], where deviations from expected anatomy are identified by measuring the difference between reconstructed and observed tissues. Building on top of the same intuition, Zhang, Zhu, and Willke [364] additionally use the symmetry properties of the brain to detect anomalies, while Jiang et al. [135] examine the difference in the attention masks when querying for healthy and pathological tissue in a zero-shot approach. While anomaly detection effectively leverages anatomical properties via reconstruction, a key limitation is that it only models deviations from expected anatomies, which may or may not be pathologies, but also could be anatomical variations.

Related to reconstruction-based approaches are pretraining strategies that mask regions guided by anatomical labels [320, 368], thereby increasing the number of masked patches in the regions of interest.

Alternative approaches employ multi-stage frameworks that utilize segmentation [66, 253, 330] or detection [226] results as positional priors, or incorporate learned positional priors [63], to enhance tumor segmentation performance. However, these methods typically target only a limited set of tumor types within specific anatomical regions, such as pancreatic [63, 253] or prostate [66, 330] tumors. Complementary research directions include text-guided approaches using VLM architectures for diagnosis classification [78] and cross-attention mechanisms that integrate medical reports for pancreatic tumor segmentation [64].

While most previous approaches focus exclusively on single-modality data (CT or MRI), recent works [2, 212] demonstrate that in PET/CT imaging, the CT domain provides the primary anatomical information, which can be effectively fused with the PET domain's tumor-specific data using modality-specific encoders. In competitive benchmarks such as the AutoPET[1] challenge, approaches have been developed concurrently to ours that explicitly incorporate anatomical masks within the CT domain to enhance pathology segmentation performance [141, 229, 266]. While concurrent work [141, 266] has similarly recognized the importance of leveraging anatomical knowledge for pathology segmentation, our approach extends beyond the competition-specific strategies of Murugesan et al. [229] by exploring diverse incorporation methods across multiple tasks (segmentation and detection) and imaging modalities (whole-body PET/CT and chest X-ray).

Our contribution:
We introduce APEx: an anatomy-pathology segmentation model that systematically integrates anatomical and pathological information through a novel query-based architecture, addressing the limitation that existing pathology segmentation models lack a holistic understanding of whole-body structures. Our key contributions are: (1) Novel Query-Based Joint Architecture: We develop APEx with shared pixel embeddings and an asymmetric information flow, where anatomical queries inform pathology predictions through a query mixing strategy, enabling anatomy-guided pathology segmentation that mirrors clinical workflow. (2) Systematic Integration Strategy Analysis: We conduct comprehensive ablations, testing multiple anatomy incorporation strategies, and identify beneficial architectural choices for anatomy-pathology information exchange. (3) Cross-Domain Validation: We evaluate across two domains (whole-body FDG-PET/CT and chest X-ray images) and two tasks (semantic and instance segmentation), achieving +2.0% and +3.3% improvements over strong baselines while providing insights into which anatomical

[1] https://autopet.grand-challenge.org

structures are most relevant for pathology detection. (4) Plug-and-Play Framework Extension: Building on the insights of APEx, we develop GRASP, a modular framework that leverages existing frozen anatomy segmentation models through feature alignment and pseudo-label integration, eliminating the need for auxiliary anatomical training while maintaining the anatomy-pathology integration benefits across diverse architectures.

Our work shifts anatomy-informed pathology segmentation from a narrow, disease-specific approach to a flexible, generalist approach, enabling improved detection of diverse pathologies, including tumors and foreign bodies, across multiple medical imaging domains and tasks.

2.5 EVALUATION OF SEGMENTATION MODELS

Model evaluation is essential for two main reasons. First, model training typically optimizes proxy losses rather than target metrics. The cross-entropy loss, for example, penalizes distributional misalignment and low confidence predictions, and can be applied to segmentation by treating it as per-voxel classification. However, this creates a gap between what we optimize and what we actually care about. Second, training performance provides no guarantee of generalization to unseen data, making independent evaluation necessary.

Inspired by our preceding work [128], we observe that a large number of works treat the task of identifying tumors in CT or PET/CT scans as a semantic segmentation task [5, 10, 73, 80, 81, 106, 107, 241, 371]. We previously discussed datasets related to this body of work in Section 2.1.

While there are works [87, 108, 142, 143, 352, 353, 375] that treat tumor identification as an object detection task, they operate on 2D or 2.5D slices. The strong performances of semantic segmentation models across different domains [120], and the easier optimization task of volumetric semantic segmentation compared to volumetric object detection [21, 125] due to the pixel-wise learning setup, have led to the widespread formulation of volumetric tumor localization as a semantic segmentation task.

Naturally, semantic segmentation models are commonly evaluated using semantic segmentation metrics, such as the Sørensen-Dice coefficient [61], Intersection over Union (IoU) [124] or Normalized Surface Dice [235, 281]. These metrics share a common limitation: They only consider a simple overlap between the predictions and the target without considering individual instances. Furthermore, these metrics are heavily biased towards large instances, as they naturally account for the majority of the volume. The resulting inflated importance of large metastases, however, does not necessarily reflect their clinical relevance. Smaller lesions can be equally or more significant for patient prognosis and treatment planning than large lesions. Current medical guidelines, including the gold standard TNM staging system, demonstrate that metastatic lesions as small as 0.2mm are clinically meaningful when they represent spread to novel organ sites, automatically upgrading patients to Stage IV disease regardless of lesion size. Consequently, evaluation metrics that inherently downweight small lesions may not accurately reflect the clinical utility of tumor detection systems, potentially leading to models that perform well on traditional segmentation metrics but miss clinically critical small metastases.

While a natural approach would be to use instance segmentation metrics such as mean average precision (mAP) [67, 183], these metrics require network predictions to have an explicit object notion, which semantic segmentation models inherently lack.

To address this issue, several metrics have been developed to work with semantic segmentation outputs while still distinguishing between the segmentation of individual objects. The most well-known work is the Panoptic Quality (PQ) metric [153], which assigns a prediction to a ground-truth if the IoU between the masks is at least 50%. Via the identification of True Positives (TP), False Positives (FP), and False Negatives (FN), these counting-based metrics are directly combined with overlap-based metrics such as IoU, thereby aiming to strike a balance between size-agnostic counting-based metrics and size-sensitive overlap measures that account for segmentation quality. The threshold of 50% to match predictions to ground-truth components generates a unique matching from predicted segments to ground-truth segments, but it also imposes a fixed threshold that can lead to sudden changes in the behavior of the metric, which we explore in Chapter 5.

Concurrent to ours, two lesion-aware dice modifications have been proposed: Moawad et al. [225] introduced lesion-dice in the context of brain-tumor segmentation, which builds upon the insights of Panoptic Quality but no longer enforces a unique matching and allows multiple predictions to be assigned to multiple ground-truths. While this adds flexibility, it introduces certain drawbacks from the non-unique mapping of predictions and ground-truth segments. [271] introduces ccDice, a topology-aware metric that generalizes the Dice score from pixel-level to connected component-level evaluation. ccDice establishes bijective mappings between predicted and ground-truth components using embedding scores with a configurable overlap threshold and a greedy heuristic to ensure a unique one-to-one matching.

Our contribution:
We introduce CC-Metrics: A novel evaluation protocol for semantic segmentation in multi-instance detection scenarios where each connected component matters equally. Our key contributions are: (1) Generalized Voronoi Partitioning: We develop a proximity-based spatial partitioning approach that assigns each pixel to its nearest ground-truth component, enabling threshold-free matching without arbitrary overlap requirements. (2) Per-Component Metric Evaluation: We demonstrate how existing segmentation metrics (Dice, Surface Dice, Hausdorff Distance) can be computed locally within each Voronoi region, providing equal weight to all components regardless of size while maintaining metric interpretability. (3) Com-

prehensive Evaluation: We show through extensive simulations and real model evaluations on PET/CT datasets that CC-Metrics reveals clinically relevant performance differences masked by traditional global metrics, particularly for scenarios involving missed small lesions. Our work addresses the critical gap between semantic segmentation evaluation and clinical requirements in multi-instance medical scenarios, providing more informative assessment tools that align with clinical priorities in AI-assisted diagnosis.

Part II

ANATOMICAL KNOWLEDGE WITHIN THE MODEL DEVELOPMENT LIFECYCLE

3 AUTOMATED DATASET GENERATION

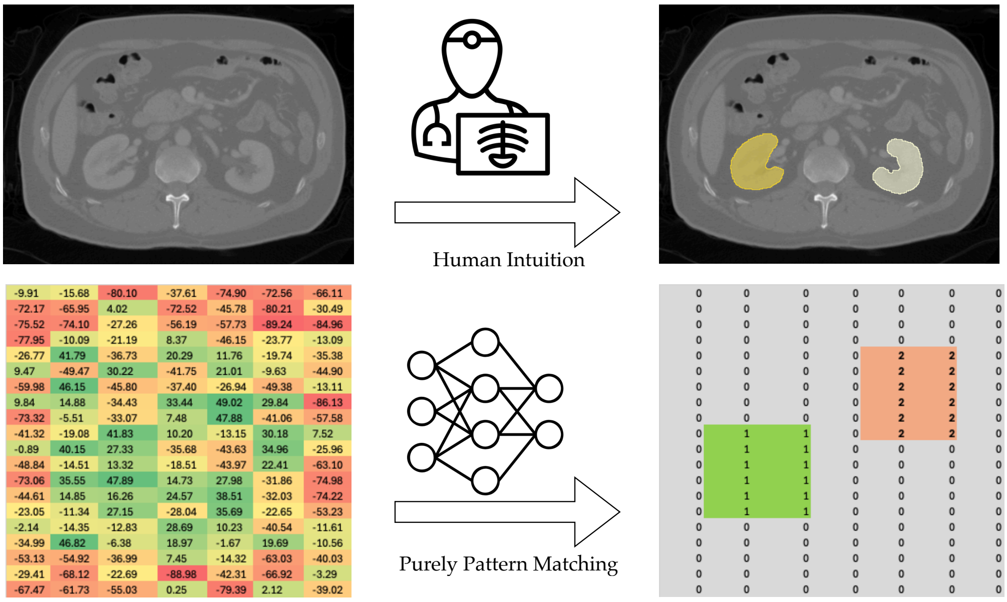
A key promise of the deep learning paradigm is the automated extraction of useful features purely from data, as opposed to model-driven approaches that require a careful selection of model parameters. This necessitates a sufficiently large amount of data, which serves as the source of information for fitting the models' parameters. Within the medical field, this poses a problem due to the required expertise to create such datasets. Although physicians can annotate a limited set of images, this approach becomes progressively impractical as data demands increase. Within this Chapter, we first take a look in Section 3.1 on how to create a large-scale dataset in an automated fashion, which is based on our ICIP 2024 contribution [130]. We then turn towards a data-centric discussion on the impact of potential errors in such large-scale datasets in Section 3.2 (currently in submission [127]) and their impact on model performances within training and pretraining settings. While understanding label quality is crucial, segmentation models can still generate non-satisfactory masks regardless of training label quality—whether due to faulty training labels or domain shifts during inference. To address the practical challenge of imperfect predictions in clinical workflows, we introduce a dynamic adaptation pipeline that allows physicians to interactively correct label predictions in Section 3.3, based on our ISBI 2025 publication [105].

There are numerous technical methods for storing images and their associated labels, including DICOM[1], NRRD[2], and NIFTY[3] file formats. These formats are primarily designed to provide pixel-wise or, in the case of volumetric imaging modalities such as CT or MRI, voxel-wise information that maps spatial image locations to corresponding semantic labels. This detailed spatial information is crucial for training deep learning models, as algorithmic approaches interpret image data merely as arrays of numerical values. An intuition is provided in Figure 3. While deep learning models were initially developed within the natural image domain [58, 163, 197, 293], they have taken the medical field by storm [53, 120, 219, 268] and have become the de facto standard for a wide variety of tasks. Crucially, these models require annotated data from which they can infer the desired mapping function from image to label space.

[1] https://www.dicomstandard.org/

[2] https://teem.sourceforge.net/nrrd/index.html

[3] https://nifti.nimh.nih.gov/



-9.91	-15.68	-80.10	-37.61	-74.90	-72.56	-66.11
-72.17	-65.95	4.02	-72.52	-45.78	-80.21	-30.49
-75.52	-74.10	-27.26	-56.19	-57.73	-89.24	-84.96
-77.95	-10.09	-21.19	8.37	-46.15	-23.77	-13.09
-26.77	41.79	-36.73	20.29	11.76	-19.74	-35.38
9.47	-49.47	30.22	-41.75	21.01	-9.63	-44.90
-59.98	46.15	-45.80	-37.40	-26.94	-49.38	-13.11
9.84	14.88	-34.43	33.44	49.02	29.84	-86.13
-73.32	-5.51	-33.07	7.48	47.88	-41.06	-57.58
-41.32	-19.08	41.83	10.20	-13.15	30.18	7.52
-0.89	40.15	27.33	-35.68	-43.63	34.96	-25.96
-48.84	-14.51	13.32	-18.51	-43.97	22.41	-63.10
-73.06	35.55	47.89	14.73	27.98	-31.86	-74.98
-44.61	14.85	16.26	24.57	38.51	-32.03	-74.22
-23.05	-11.34	27.15	-28.04	35.69	-22.65	-53.23
-2.14	-14.35	-12.83	28.69	10.23	-40.54	-11.61
-34.99	46.82	-6.38	18.97	-1.67	19.69	-10.56
-53.13	-54.92	-36.99	7.45	-14.32	-63.03	-40.03
-29.41	-68.12	-22.69	-88.98	-42.31	-66.92	-3.29
-67.47	-61.73	-55.03	0.25	-79.39	2.12	-39.02

0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	2	2	0
0	0	0	0	2	2	0
0	0	0	0	2	2	0
0	0	0	0	2	2	0
0	0	0	0	2	2	0
0	1	1	0	2	2	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
0	1	1	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Figure 3: Conceptual comparison between human (radiological) image interpretation (top) and model-driven segmentation (bottom). Radiologists rely on semantic understanding and anatomical priors, whereas models are parameterized functions trained to map pixel- or voxel-wise intensity values to labels. These models adjust their weights purely based on data, without intrinsic anatomical awareness.

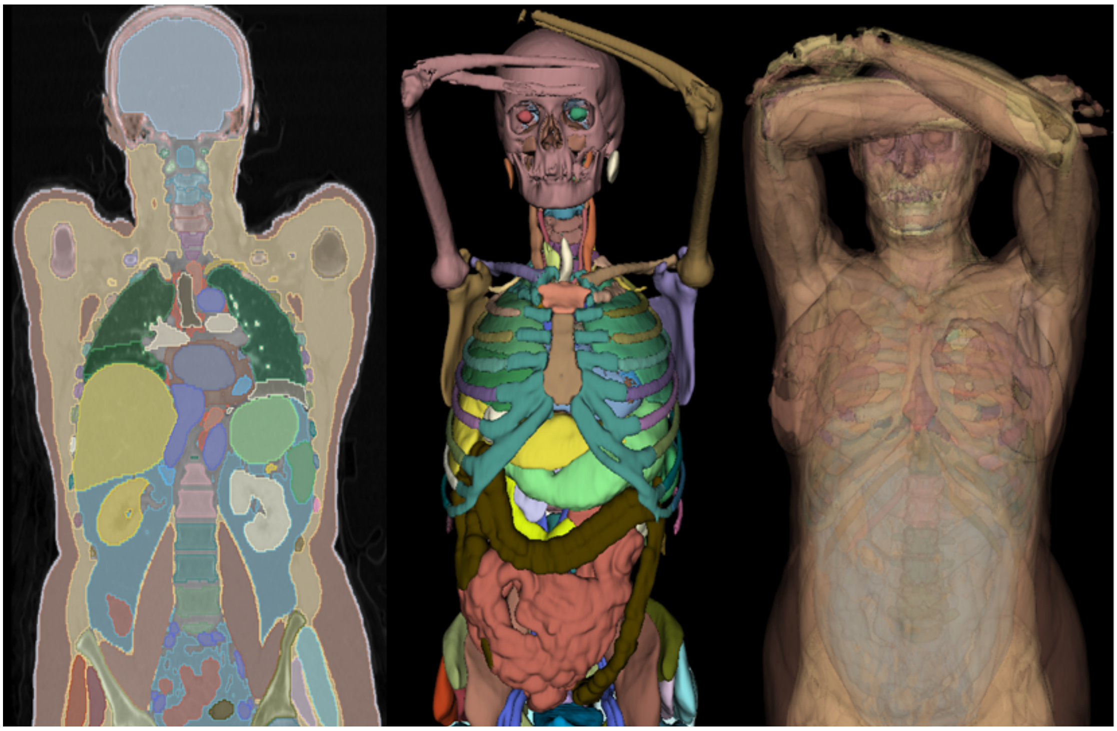


Figure 4: A sample image of the proposed Dense Anatomical Prediction Atlas dataset in a coronal slice next to two 3D images with and without soft tissue rendered by 3DSlicer [151].

3.1 AUTOMATED DATASET CONSTRUCTION OF DAP-ATLAS

This section is based on our publication in ICIP 2024 [130]

Providing high-quality annotations in the medical field typically requires the annotator to have undergone in-depth medical training (e.g., becoming a radiologist), which makes the generation of large datasets by scaling dataset annotations to thousands of cheap annotators not feasible. Importantly, medical professionals must carefully balance the allocation of time between patient care and auxiliary tasks; dedicating substantial effort to data annotation is often not feasible. Moreover, to obtain gold-standard annotations, multiple expert segmentations should be collected and merged using consensus-based approaches such as the STAPLE [332] algorithm to mitigate individual bias and ensure reliability. For example, the recently proposed ATM-dataset [367], which is comparable in size to the proposed Dense Anatomical Prediction (DAP) Atlas dataset, utilizes three expert radiologists working on a single label, with each CT volume requiring approximately 60-90 minutes [367]. Assuming an average annotation time of 75 minutes per CT volume for this fine-grained class, creating this single-label dataset would require approximately 80 full working days, based on an 8-hour uninterrupted work schedule per day. Scaling this process to encompass multiple labels, even considering reduced annotation times for simpler structures, becomes clearly impractical. This is not only due to the significant increase in workload but also due to the heightened risk of radiologists' fatigue and potential errors arising from the increased complexity of the task.

Given these substantial annotation challenges and resource constraints, it is unsurprising that current models for medical data predominantly specialize in partially annotated datasets on sub-areas of the human body. These datasets range from single-organ annotations, such as the segmentation of the spleen or pancreas [294], to multi-organ datasets, such as the BTCV [168], which contains annotations for 13 abdominal organs. We provide an extensive overview of these datasets in Section 2.1.

The recently introduced Medical Segmentation Decathlon [5] aims at generalizing models to multiple tasks. Each of the individual tasks, however, remains limited to a specific body region and a particular organ of interest, which does not provide a comprehensive anatomical view.

Other works [300, 333] that aim to create large-scale datasets have adopted human-in-the-loop strategies. The creators of MOOSE [300], a multi-organ segmentation model, employ a hybrid approach that combines expert annotation and automated annotations. The experts annotated a small dataset of 50 CT images, comprising 13 organ structures, 20 bone segments, and 4 tissue structures, which were semi-automatically extracted. For the majority of their labels, which are 83 cerebral structures, the authors use an automatic segmentation approach by leveraging the Hammersmith atlas [94].

TotalSegmentator [333] approaches this problem via a human-in-the-loop strategy in which an expert improves model predictions, which are again fed to retrain the model to improve its predictions. While this procedure noticeably reduces the time an expert spends on annotation, it still relies on direct interaction with experts for multiple weeks to generate a dataset of 104 anatomical structures. It is also necessary to acknowledge that large datasets with this many labels present a significant challenge when evaluating label quality, as it becomes increasingly infeasible to manually verify pixel-wise alignments for all anatomical structures. As such, TotalSegmentor, in its initial version [335], relied on quality assurance via 3D renderings instead of manual voxel-wise alignment checks.

When comparing the volumetric field to the two-dimensional image domain, several works utilize entirely automated annotation for classification and segmentation purposes [154, 280, 345] in both the medical and natural image domains. One example is the recent *Segment Anything*-dataset [154]. Here, the authors train a base model on manually annotated data and make predictions on unlabeled images through multi-scale inference and filter predictions via non-maximum suppression, leading to 11 million annotated images. The general concept of pseudo-label filtering from weak annotations like image-level class labels improves the unlabeled training data and leads to more stable models[223, 255, 263, 280, 321]. Other works [115] have demonstrated the successful application of pseudo-generated tumor labels on CT data of the liver, resulting in accurate segmentation results for real liver tumors.

We build upon the core idea of pseudo-label filtering and refinement to employ an expert-free dataset generation approach that aggregates the scattered anatomical knowledge of multiple source datasets into a single whole-body target dataset. We combine anatomical information from various sources through pseudo-label-based label aggregation and pseudo-label refinement via post-processing strategies, leveraging anatomical textbook knowledge to verify the anatomical plausibility of the labels.

Through this strategy, we develop the Dense Anatomical Prediction Atlas dataset (DAP Atlas), which aggregates the scattered anatomical knowledge from multiple source datasets into a single dataset via sophisticated pseudo-label refinement through post-processing strategies that leverage anatomical textbook knowledge to assert the plausibility of the labels. This dataset is the first to contain dense annotations for almost every voxel in a full-body human CT.

The DAP Atlas has been approved by medical experts, despite not having been manually annotated for its creation. The dataset consists of 533 whole-body CT images with labels for 142 anatomical structures, ranging from general body composition tissues and organs to bones and various vessels.

We believe that models trained on the DAP Atlas may serve a variety of clinically relevant downstream tasks that benefit from extensive anatomical knowledge, such as body composition analysis, surgery planning, or cancer treatment monitoring.

3.1.1 *Data Acquisition*

We begin by selecting a foundational dataset for label aggregation. Naturally, we choose a whole-body dataset that allows us to collect labels from the entire anatomy. The recently published AutoPET dataset [81] offers an extensive collection of whole-body PET/CT scans pertinent to therapy monitoring in nuclear medicine.

We select suitable scans from the dataset based on consistent slice thickness in the axial plane, ensuring a homogeneous dataset that encompasses the body at least from the head down to slightly below the hips. This approach allows the delineation of anatomical structures across the head-neck, thorax, abdominal, and pelvic regions. We show two examples that met the selection criteria in Figure 5.

DAP Atlas exhibits age, gender, and pathology distribution characteristics comparable to those of the AutoPET dataset. A detailed descriptive analysis of these demographic and clinical dimensions is provided in Figure 35 in the appendix.

3.1.2 *Knowledge Aquisition*

The DAP Atlas dataset aggregates multiple sources of anatomical segmentation knowledge, which we categorize into public knowledge, represented by publicly avail-

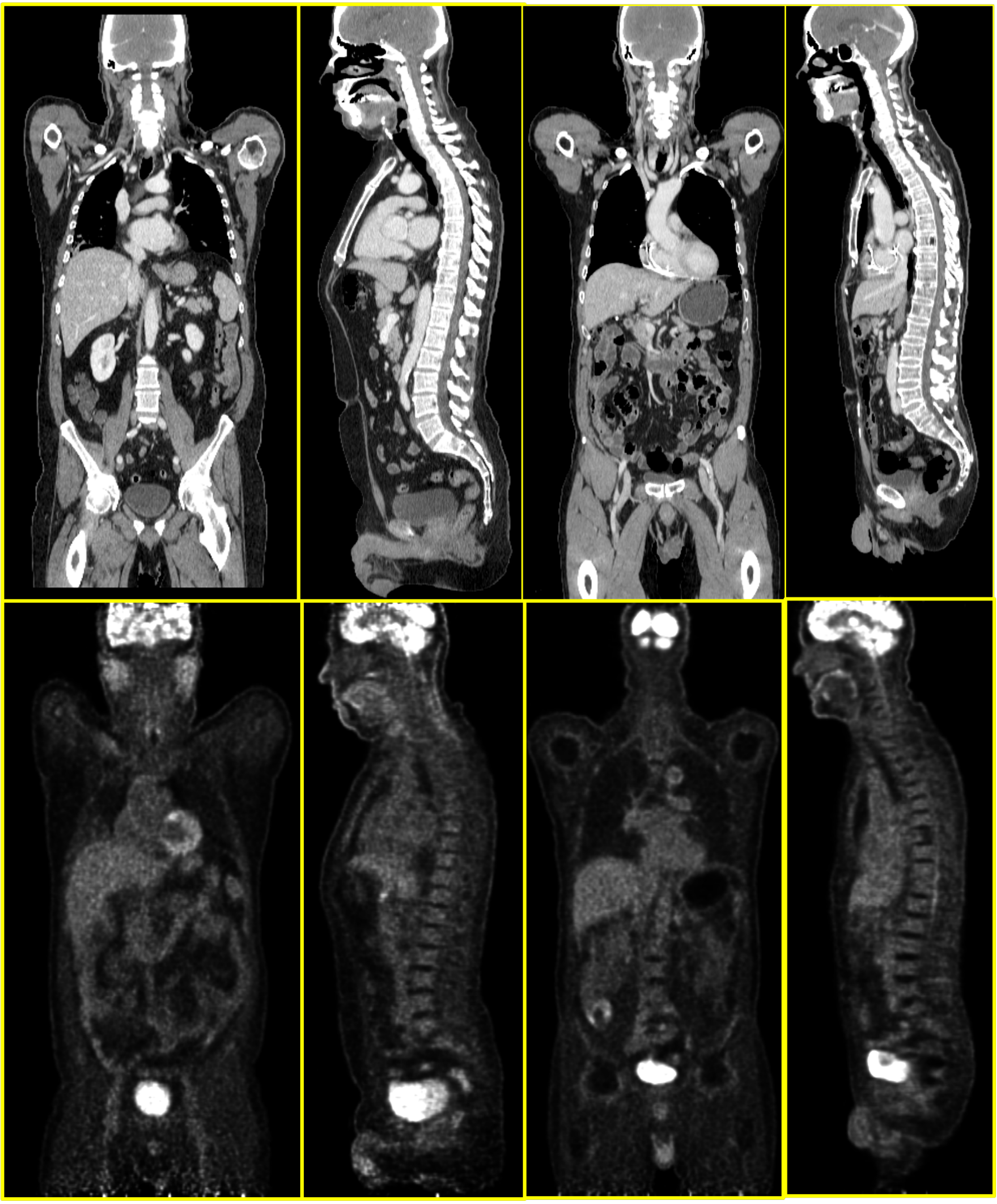


Figure 5: Illustrated are two CT scans (top row) in coronal and sagittal views, selected from the AutoPET dataset to serve as the foundation for the DAP Atlas dataset. As specified, each scan encompasses anatomical regions from the head and neck down to below the pelvis. The bottom row presents the corresponding PET scans for these patients.

able datasets, and private expertise, comprising private datasets accessible to us. We consider eleven public datasets, namely Pediatric [139], Total Segmentator [335], SegThor [167], CT50Abdomen [206], MAL Cervix [168], Amos [133], RibSeg [355], Verse [282], ATM [367], PARSE [201], Pelvic CT [187], which span segmentations of differing origins and structures. We describe these datasets in detail in Appendix C.1.1 in the appendix.

We extract the knowledge represented through the labels by using ensembles of nnU-Nets [120], a framework that automatically configures a U-Net [268] to learn the labels of the respective datasets. The publicly available datasets employed are listed alongside their corresponding label contributions to the DAP Atlas in Figure 7. After training, we predict the learned labels into the selected AutoPET subset.

In addition to the previously described public datasets, we utilize non-publicly available datasets and models to incorporate unique and previously unavailable labels. One of the models is the body composition analysis model [156], which differentiates between different types of tissue, such as *fat* or *muscles*. In total, we extract 9 labels from the body composition model source. Furthermore, we utilize a private dataset source consisting of 104 diverse head and neck contrast CT images from four different source cohorts [22, 23, 54, 86, 296, 315] to add 12 unique, previously unavailable labels in the head-neck region, such as the *artery subclavian*.

After obtaining the labels from the different nnU-Net predictions, we use anatomically derived rules to refine the current predictions and generate 7 additional labels. An intuitive example for a new label that can be derived from the combination of obtained labels and medical common sense is the *skull*. It can be derived from a thresholding procedure obtained by the bone window present in CT images. Bones typically yield CT values between 300 and 3000 Hounsfield Units (HU), which serve as the described thresholds. The obtained set of voxels can be restricted to the area above the C5 vertebra, which was previously obtained. Finally, we remove already predicted vertebrae from the thresholded voxels, which leaves us with an accurate mask for the skull. We also exploit the behavior of the neural network predictions, which have been trained only on parts of the anatomy and typically confuse structures that appear similar in the CT images. Common systematic errors include predicting the gonads as the eyeballs or the colon as the nasal cavity due to the presence of air in these areas. We exploit these systematic mistakes and remap the produced labels according to their location within the human body. By employing these simple rules, we add 7 additional labels

3.1.3 *Knowledge Aggregation*

Label Merging:

The workflow for constructing the DAP Atlas is illustrated in Figure 6. To aggregate the predictions of the individual models, we define a common labeling scheme to which we map the obtained masks. When integrating the different anatomical structures into the Atlas labeling scheme, we aggregate them according to their anatomical hierarchical level from coarse to fine, starting from abstract tissue classes such as *muscles* or *fat*. We gradually add the different organs and finally fine-grained vessel structures such as *Pulmonary Arteries*. During the aggregation process, we employ basic anatomical rules to improve individual predictions. Since several labels represent multiple versions of the same anatomical structure, such as the class *aorta* found in the source datasets Total Segmentator, Amos, and SegThor, it becomes essential to integrate these predictions. Typically, we achieve this by merging the predictions from models trained on different source datasets into a unified mask, representing the union of all individual masks. For instance, the SegThor [167] model predicts the *aorta* only within the thoracic region, despite the aorta extending beyond this boundary. By combining this partial mask with additional masks generated by other models, we obtain a comprehensive mask that accurately represents the aorta across the entire anatomy. This integration process consolidates the anatomical knowledge from diverse datasets into a singular, cohesive representation, which constitutes the primary objective of this work.

Self-Training: After integrating the heterogeneous labels into the DAP Atlas label set, which includes the rule-based novel labels, we generate a single, seamless dataset and perform one iteration of self-training. The benefits of this procedure are fourfold: (1) It unifies the heterogeneous label resolution stemming from the original datasets. The expert models typically resample the target image to the resolution at which they were trained. (2) It eliminates non-systematic random noise due to the network receiving consistent feedback only from consistent predictions, a phonemenon well known [189, 190], (3) It distills the fragmented knowledge into a single model capable of predicting the entire anatomy, thus massively decreasing the necessary inference time, and lastly, (4) Self-training hampers the exact reconstruction of private data from expert models which were directly trained on private source datasets.

We generate the raw version of the DAP Atlas dataset by applying the obtained unified anatomical model to the selected DAP Atlas target volumes. While the overall label quality is good, we observe systematic errors where the networks repeatedly mislabeled voxels of paired structures—for example, voxels of the right kidney were included in the mask of the left kidney and vice versa, or vertebrae were confused with their adjacent counterparts. Furthermore, we observe implausible predictions of structures within body regions that are not possible, e.g., the colon being predicted outside the abdomen. Finally, we observe structures belonging to the reproductive system to be predicted for the wrong sex. These errors can be corrected by reapplying anatomical rules, which we formalize in Algorithm 3 and detail in Appendix C. We employ two additional rule sets to filter implausible predictions: (i) during rib counting (Algorithm 7), we exclude predictions that produce inconsistent rib orderings when derived from median versus minimum rib points; (ii) during left–right splitting (Algorithm 8), we discard predictions where the hyperplane defined by vertebral centroids deviates excessively from the expected axial orientation. An exemplary improvement through the developed post-processing algorithm is displayed in Figure 8.

After applying Algorithm 3 to the raw labels, we receive the final version of the dataset, which is rated as very impressive by a consulted radiologist. We describe the extensive validation procedure of the dataset in Section 3.1.6.

We further develop a prediction model that relies less on algorithmic label improvement and can be used to predict labels directly present in our DAP Atlas dataset. We describe the procedure in detail in Appendix C.1.4. In the following, we will refer to the first version of the model, which was previously described as the *Atlas dataset model* (V1), as this is the one that was used for the dataset creation, and to the prediction model as the *Atlas prediction model* (V2).

3.1.4 *Data Access*

We make the code for the dataset aggregation, trained models, post-processing scripts, and the dataset itself publicly available[4]. The dataset is also hosted on synapse.org/#!Synapse:syn52287632.1 under a CC-BY 4.0 license, ensuring long-term availability.

DAP Atlas builds on top of the AutoPET dataset [81], which can be accessed on The Cancer Imaging Archive (TCIA) under its collection name “FDG-PET-CT-Lesions” to download the raw PET/CT data. We have retained the AutoPET naming convention in DAP Atlas to ensure that the masks can be easily matched with their corresponding original CT volume.

```
DAP Atlas Anatomical Labels
 ├── AutoPET_0011f3deaf_10445.nii.gz
 ├── AutoPET_01140d52d8_56839.nii.gz
 ├── AutoPET_0143bab87a_33529.nii.gz
 └── ...
```

The given name consists of the subject ID followed by the last 5 digits of the Study UID, which allows a unique matching of the segmentation masks to the AutoPET CTs.

3.1.5 *Compliance with Ethical Standards*

Originally, for the CT source data from the AutoPET dataset, the ethics committee of the University Hospital Tübingen waived the necessity for ethical approval for anonymized publication data. Further ethical approval was not required, as confirmed

[4] https://github.com/alexanderjaus/AtlasDataset

by the license attached with the open-access data (CC-BY 4.0) as confirmed by the authors [81].

For public label sources, a public competition ethical approval was not required, as confirmed by the license attached with the open-access data. For private label sources, either informed consent was obtained from all patients for use of anonymized data in retrospective studies [296] and/or Institutional Review Board approval was obtained [156, 296].

3.1.6 *Technical Validation*

As previously discussed, we propose the DAP Atlas dataset as a knowledge aggregation dataset from multiple fragmented source datasets, which are impractical to train neural networks on, as they only offer partial supervision for the presented anatomical structure and label everything else as background. As the DAP Atlas dataset consists of numerous volumes and is rich in labels, it is nearly impossible for experts to verify the correctness of every voxel. Other datasets containing few annotations can still use manual label checking and correction. The Airway Tree Modeling dataset [367], which is comparable in its number of CTs, provides annotations for a single label. As previously stated, its creation time was 80 radiologist days. This demonstrates that even manually checking and correcting labels in DAP Atlas at the voxel level is nearly impossible. TotalSegmentator [335] accelerates verification by utilizing 2D renderings of 3D organs; however, this approach still cannot guarantee voxel-wise correctness.

3.1.6.1 *Evaluation Setup:*

To address the aforementioned problem of evaluation, we propose a hybrid approach that combines human experts, anatomical plausibility, and usefulness for the Deep Learning community.

- **Deep Learning Applicability:** We verify the usefulness of our dataset for the development of Deep Learning algorithms by taking our anatomical segmentation models to the test on the BTCV [168] abdomen dataset. This dataset was not used in the construction of the dataset and provides an unbiased performance check. We compare the performance of the Atlas dataset model and the Atlas prediction model in Table 3

- **Expert Checks:** To obtain human feedback while accounting for the limited availability of medical experts, we asked a radiologist to evaluate a subsample of 25 randomly selected volumes from the DAP Atlas dataset.

- **Anatomical Insights of DAP Atlas:** To verify the general, global anatomical plausibility of the dataset, we use the labels of the DAP Atlas dataset to calculate the volumes and mean intensities as characteristic descriptors of the different anatomical structures. We plot these descriptors against the age and gender of the patients and verify if they follow characteristic medical curves. Furthermore, we compare the volume distributions of Atlas organ masks with several annotated datasets to investigate deviations in volume distributions across different datasets. Finally, we investigate which anatomical structures in the Atlas dataset are most affected by which type of cancer.

By combining these three evaluation approaches, we leverage the thoroughness of human expert local voxel-wise checks with the scalability of global overall checks, ensuring that the dataset introduces merit to the Deep Learning community.

3.1.6.2 *Evaluation Results:*

In the following sections, we present the results of applying the established evaluation protocol to our dataset.

Deep Learning Applicability

Regarding the usefulness of the provided dataset for Deep Learning models, we test the developed Atlas models on the BTCV [168] dataset. This dataset has not been used to construct the DAP Atlas dataset and provides an unbiased performance check. We emphasize that we do not fine-tune the models using the BTCV training data, but perform inference on the BTCV test set without post-processing. We report the performance measures of the Atlas dataset model and the Atlas prediction model in Table 3.

Our Atlas dataset model achieves an average Dice score of approximately 81%. The Atlas prediction model (V2), developed through iterative training and post-processing cycles, achieves an 85% dice score, an improvement of ∼ 4%, demonstrating its increased robustness due to the adapted training schedule. 85% dice is on par with state-of-the-art medical segmentation models such as UNETR [99], which are trained in a standard supervised fashion on the training dataset. The Atlas prediction (V2) model shows significant improvements in abdominal structures, i.e. 81% vs 85% for the *vena cava inferior (IVC)* or 75% vs 83% for the pancreas, and in particular smaller structures such as *Adrenal Glands*. But we also notice small decreases in performance for *Left and Right Kidneys*. Regarding the Mean Surface Distance performance, we observe an overall improvement in the Atlas prediction model compared to the Atlas dataset model; however, the improvements in individual organs do not follow a clear pattern. In summary, we find that the DAP Atlas dataset enables the creation of high-quality anatomical models, capable of delivering predictions on par with those of models trained on the respective datasets.

A qualitative assessment of the label quality is shown in Figure 9. Our model predictions exhibit a remarkable level of alignment with the ground truth. Beyond that, we also show how the BTCV organs are a well-integrated small fraction of the anatomical structures of the human body, which our model is able to segment.

	Spl	RKid	LKid	GB	Eso	Liv	Sto	Aor	IVC	PV&SV	Pan	RAd	LAd	Tot
Dice Scores (%)														
Atlas (V1)	96	91	94	62	72	97	84	91	81	76	75	64	59	81
Atlas (V2)	96	86	92	72	80	96	86	92	85	77	83	75	74	85
Mean Surface Distance														
Atlas (V1)	1.14	2.21	0.83	–	1.68	0.79	3.26	1.51	2.03	1.48	2.53	1.30	1.67	2.03
Atlas (V2)	0.65	4.28	1.70	–	1.38	1.21	2.50	1.29	1.87	1.77	1.51	0.80	0.90	1.85

Table 3: Class-wise Dice and Mean Surface Distance performance on the BTCV dataset for the Atlas dataset (V1) and Atlas prediction model (V2). Predictions of both models are evaluated without post-processing. Both models deliver convincing performances; however, the robust Atlas V2 model benefited from the adapted training procedure. Abbreviations: Spl=Spleen, RKid=Right Kidney, LKid=Left Kidney, GB=Gallbladder, Eso=Esophagus, Liv=Liver, Sto=Stomach, Aor=Aorta, PV&SV=Portal Vein & Splenic Vein, Pan=Pancreas, RAd=Right Adrenal, LAd=Left Adrenal, Tot=Total.

Expert Checks:

We include a human expert in the quality check pipeline. To gather human feedback on the DAP Atlas dataset, we randomly sampled 25 volumes and had an expert radiologist evaluate the quality, discuss shortcomings, and explore applications. The following section discusses the feedback we received and displays the strengths and weaknesses of the DAP Atlas dataset.

Overall, the feedback that we received was mostly positive:

> *Overall, for whole-body segmentation of a normal patient, it's very impressive. [...] It was also good on some patients pointing out a small hiatal hernia. Otherwise, I think it is more useful for medical students and internal medicine doctors who may not be as familiar with anatomy on CT.*

In addition to this general feedback, we gained some insights into the structural mechanics of the dataset, which we summarize below. The expert noted that some structures do not always seem to be homogeneously segmented, naming predominantly the spinal canal. Further, for tree-like structures such as the pulmonary artery, the fine-grained branch endings lose detail and become under-segmented. Lastly, it was noted that the borders of abutting abdominal organs are at times offset and differ from the expert's estimation.

Anatomical Insights of DAP Atlas:

As a first global check, we compare the volume distributions of the masks in our proposed DAP Atlas dataset against other datasets that were annotated by experts. The different volume distributions are shown in Figure 10.

The selection of anatomical structures for which we display the distribution plots is based on the criterion that at least two additional datasets, in addition to the proposed DAP Atlas dataset, contain the structure. By observing the distributions, we find that the volume distributions for the same organs in different datasets do vary by small amounts, but the general shape of the distributions is very similar among the datasets. Small differences in the distributions of organ volumes across the datasets are quite plausible and may stem from limited sample sizes, different annotation schemes, or the patient selection criteria. For instance, the distributions of organs in the Pediatric dataset tend to be shifted to the left, which can be easily explained since the dataset focuses on patients below the age of 18 years. Larger variations and in particular distributions deviating towards the origin can stem from CT images only covering parts of organs, which is common in the Total Segmentator dataset. When comparing the DAP Atlas dataset to the family of organ distributions, we find it to be well-integrated in terms of its distribution support and shape.

To analytically confirm this, we compute the Jensen–Shannon Divergence (JSD), a symmetric, finite measure that calculates the deviations between distributions. For each of the analyzed anatomical structures, we calculate the JSD of a dataset's volume distribution to all other volume distributions within the same structure. We average these values to receive the distribution's average divergence to all other distributions for a given structure. The greater the average value, the more distant the volume distribution is from its peers. Finally, we create a box plot to compare the distribution of average divergences across datasets. We find that the DAP Atlas dataset is well-placed among the other datasets, with distributional agreements that are very similar to those of voxel-wise expert-annotated datasets.

As a second global check, we calculate the volume and mean intensity of each anatomical structure in the DAP Atlas dataset and plot them against the age of the patients in Figure 11.

We organize Figure 11 into three parts. The three scatter plots on the left side display the volume of the mentioned organ masks in milliliters plotted against the age of the patients, which can be obtained from the metadata. Two general observations are apparent: first, organ volumes in female patients are systematically smaller than in male patients. Second, quadratic fits reveal well-described medical patterns. The liver displays a downward-facing parabola, increasing during early adulthood, reaching a plateau, and declining with advanced age, consistent with known hepatocyte loss and reduced perfusion. The left atrium shows a monotonic increase in volume, reflecting the cumulative effect of reduced ventricular compliance and age-related diastolic dysfunction, both of which elevate atrial load. Both of these behaviors are well-known medical facts [149], confirming the anatomical plausibility of the dataset. The left ventricle, in contrast, follows an inverted U-shaped trajectory, peaking around the age of 50 years in our cohort before declining.

The three scatter plots on the right display the calculated mean intensities, measured in HU of the respective structures in the CT, as indicated by the corresponding Atlas masks. We choose three characteristic examples and observe physiologically consistent trends: hip bone density declines with age, reflecting osteoporosis; liver attenuation decreases, consistent with age-related fat accumulation; and finally, an interesting observation can be derived from the last plot of the skull which presents an outlier: The reason is a patient with dental implants pushing up the average HU-values of the patient's skull due to extreme hardness.

In the bottom row, we show an example of a potential future use case in which the Atlas dataset may serve as a cornerstone in the joint investigation of the entire anatomy and pathologies. We calculate which of the known structures in the Atlas dataset are most affected by which type of cancer. For each patient, we examine which anatomical structures are affected by cancer. When determining if an anatomical structure has been affected by cancer, we consider it to be cancerous if there is at least one voxel labeled as cancerous tissue. During this analysis, we do not distinguish between metastasis and primary cancer cells. Finally, we normalize by the total number of patients with the respective disease to determine the likelihood that an anatomical structure is affected, stratified by a given diagnosis.

3.1.7 *Limitations of the Dataset*

When using the dataset, certain limitations of the provided labels should be considered. First, the cohort was retrospectively collected from patients referred for oncological imaging under suspicion of lymphoma, melanoma, or lung cancer, with approximately half ultimately diagnosed with a malignancy. Thus, the data do not represent a healthy reference population, and organ morphology may be influenced by disease, treatment, or comorbid conditions. Users of the dataset may choose to filter the dataset to only include healthy subjects, thereby easing this effect.

Second, the organ volumes are derived from automated segmentations, which, despite extensive quality control, can introduce errors or occasional mislabeling. Third, the imaging protocols originate from routine clinical examinations, resulting in heterogeneity of scanner types, acquisition settings, and reconstruction parameters that may affect the consistency of the generated labels.

ID	Label	ID	Label	ID	Label
0	Background	49	Vertebra C7	97	Humerus Left
1	Unknown Tissue	50	Vertebra T1	98	Humerus Right
2	Muscles	51	Vertebra T2	99	Skull
3	Fat	52	Vertebra T3	100	Hip Left
4	Abdominal Tissue	53	Vertebra T4	101	Hip Right
5	Mediastinal Tissue	54	Vertebra T5	102	Sacrum
6	Esophagus	55	Vertebra T6	103	Femur Left
7	Stomach	56	Vertebra T7	104	Femur Right
8	Small Bowel	57	Vertebra T8	105	Heart
9	Duodenum	58	Vertebra T9	106	Heart Atrium Left
10	Colon	59	Vertebra T10	107	Heart Tissue
12	Gallbladder	60	Vertebra T11	108	Heart Atrium Right
13	Liver	61	Vertebra T12	109	Heart Myocardium
14	Pancreas	62	Vertebra L1	110	Heart Ventricle Left
15	Kidney Left	63	Vertebra L2	111	Heart Ventricle Right
16	Kidney Right	64	Vertebra L3	112	Iliac Artery Left
17	Bladder	65	Vertebra L4	113	Iliac Artery Right
18	Gonads	66	Vertebra L5	114	Aorta
19	Prostate	67	Costa 1 Left	115	Iliac Vena Left
20	Uterocervix	68	Costa 1 Right	116	Iliac Vena Right
21	Uterus	69	Costa 2 Left	117	Inferior Vena Cava
22	Breast Left	70	Costa 2 Right	118	Portal Vein and Splenic Vein
23	Breast Right	71	Costa 3 Left	119	Celiac Trunk
24	Spinal Canal	72	Costa 3 Right	120	Lung Lower Lobe Left
25	Brain	73	Costa 4 Left	121	Lung Upper Lobe Left
26	Spleen	74	Costa 4 Right	122	Lung Lower Lobe Right
27	Adrenal Gland Left	75	Costa 5 Left	123	Lung Middle Lobe Right
28	Adrenal Gland Right	76	Costa 5 Right	124	Lung Upper Lobe Right
29	Thyroid Left	77	Costa 6 Left	125	Bronchus
30	Thyroid Right	78	Costa 6 Right	126	Trachea
31	Thymus	79	Costa 7 Left	127	Pulmonary Artery
32	Gluteus Maximus Left	80	Costa 7 Right	128	Cheek Left
33	Gluteus Maximus Right	81	Costa 8 Left	129	Cheek Right
34	Gluteus Medius Left	82	Costa 8 Right	130	Eyeball Left
35	Gluteus Medius Right	83	Costa 9 Left	131	Eyeball Right
36	Gluteus Minimus Left	84	Costa 9 Right	132	Nasal Cavity
37	Gluteus Minimus Right	85	Costa 10 Left	133	Artery Common Carotid Right
38	Iliopsoas Left	86	Costa 10 Right	134	Artery Common Carotid Left
39	Iliopsoas Right	87	Costa 11 Left	135	Sternum Manubrium
40	Autochthon Left	88	Costa 11 Right	136	Artery Internal Carotid Right
41	Autochthon Right	89	Costa 12 Left	137	Artery Internal Carotid Left
42	Skin	90	Costa 12 Right	138	Internal Jugular Vein Right
43	Vertebra C1	91	Rib Cartilage	139	Internal Jugular Vein Left
44	Vertebra C2	92	Sternum Corpus	140	Artery Brachiocephalic
45	Vertebra C3	93	Clavicula Left	141	Vein Brachiocephalic Right
46	Vertebra C4	94	Clavicula Right	142	Vein Brachiocephalic Left
47	Vertebra C5	95	Scapula Left	143	Artery Subclavian Right
48	Vertebra C6	96	Scapula Right	144	Artery Subclavian Left

Table 4: Full list of available labels in the DAP Atlas dataset. Besides the self-explanatory tissues, we include a class *Unknown Tissue* indicating tissue that likely still needs to be annotated. It typically contains tissue structures that have not been annotated explicitly but were obtained by morphological operations. We still include this class as it has the potential to be useful for future work.

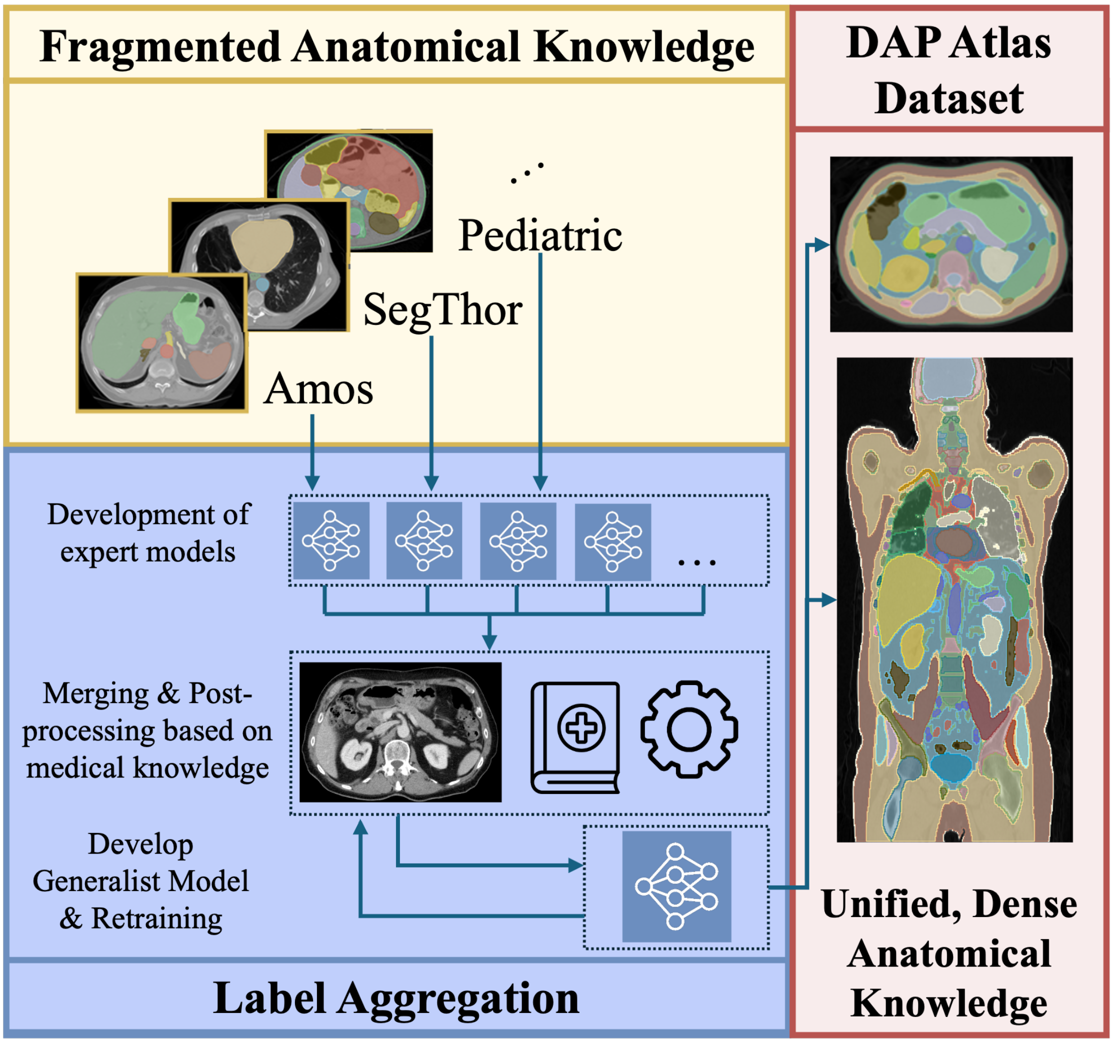
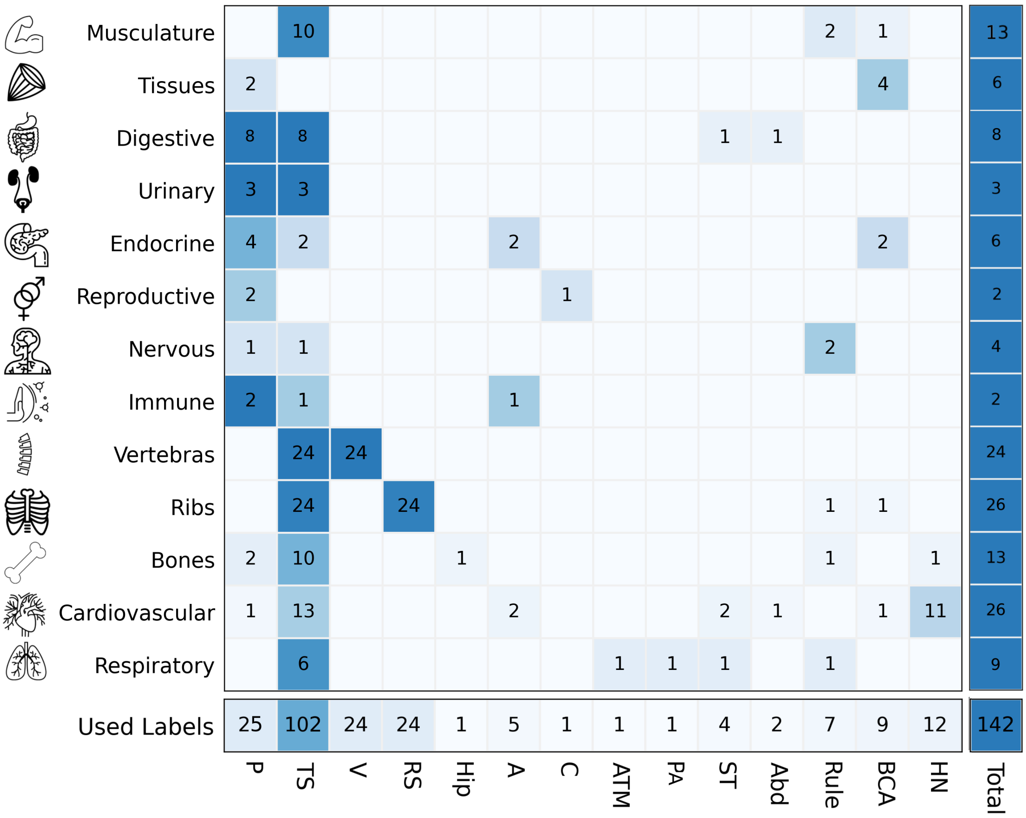

Fragmented Anatomical Knowledge
...
Pediatric
SegThor
Amos
Development of expert models
...
Merging & Post-processing based on medical knowledge
Develop Generalist Model & Retraining
Label Aggregation
DAP Atlas Dataset
Unified, Dense Anatomical Knowledge

Figure 6: The DAP Atlas dataset provides a holistic, dense label map spanning 142 anatomical labels. After training expert models on several source datasets, we aggregate their knowledge on a subset of the AutoPET dataset and refine the predictions using anatomical textbook-based rules. After one iteration of self-training, the predictions are refined by a post-processing algorithm.



	P	TS	V	RS	Hip	A	C	ATM	PA	ST	Abd	Rule	BCA	HN	Total
Musculature		10										2	1		13
Tissues	2												4		6
Digestive	8	8								1	1				8
Urinary	3	3													3
Endocrine	4	2				2							2		6
Reproductive	2						1								2
Nervous	1	1										2			4
Immune	2	1				1									2
Vertebras		24	24												24
Ribs		24		24								1	1		26
Bones	2	10			1							1		1	13
Cardiovascular	1	13				2				2	1		1	11	26
Respiratory		6						1	1	1		1			9
Used Labels	25	102	24	24	1	5	1	1	1	4	2	7	9	12	142

Figure 7: Overview of the different source datasets from which the DAP Atlas dataset is derived. On the bottom row, we show the number of labels in the DAP Atlas dataset that are influenced by the respective source dataset. The abbreviations for the datasets are as follows: **P**:Pediatric [139], **TS**: Total Segmentator [335], **V**: Verse [282], **RS**:RibSeg [355], **Hip**: Pelvis CT [187], **A**:Amos [133], **C**:MAL Cervix [168], **ATM**: ATM Airway Tree Modelling [367], **PARSE**: Pulmonary Artery Segmentation [201], **ST**:SegThor [167], **Abd**: CT50 Abdomen [206], **Rule**: Rule based derived label, **BCA**: Body composition analysis model [156], **HN**: Private Head-neck dataset.

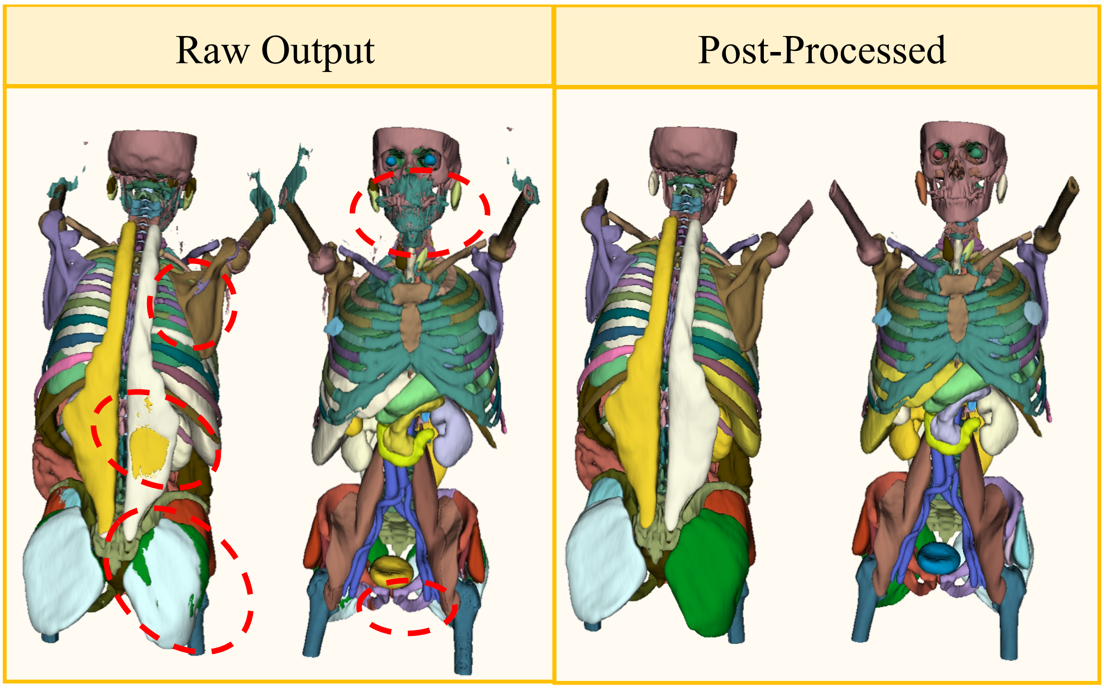


Figure 8: Comparison of the raw output of the unified model and the post-processed volume. In red, we mark problem regions in the raw labels, which are corrected by the post-processing algorithm.

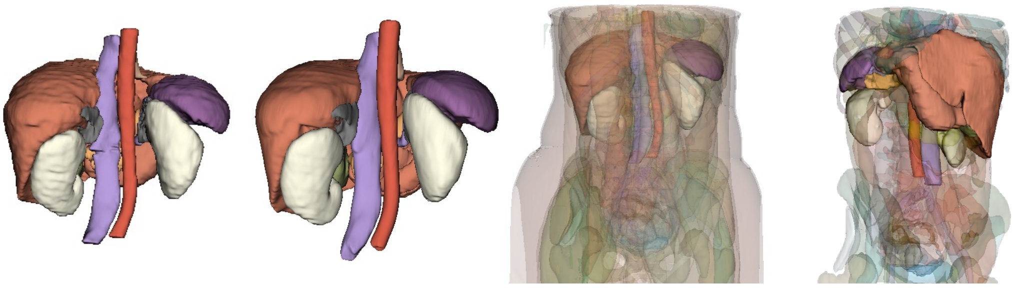


Figure 9: Qualitative assessment of the BTCV [168] out-of-distribution performance of the Atlas prediction model. As ground-truth labels for the test set are unavailable, we conduct inference on the training set for this evaluation. The model has not seen any part of this dataset, neither the training nor the test set. On the left, we show the ground truth of an example within the training dataset, alongside the predictions obtained by our model. On the right, we display all the predictions that can be obtained by our model, highlighting the 13 BTCV labels. We make two observations. The 13 BTCV organs are meticulously addressed in our approach, aligning closely with the ground truth. Furthermore, the BTCV organs are only a fraction of the structures that our model can segment. As shown on the right side, our model achieves comprehensive and dense anatomical segmentation of the human body.

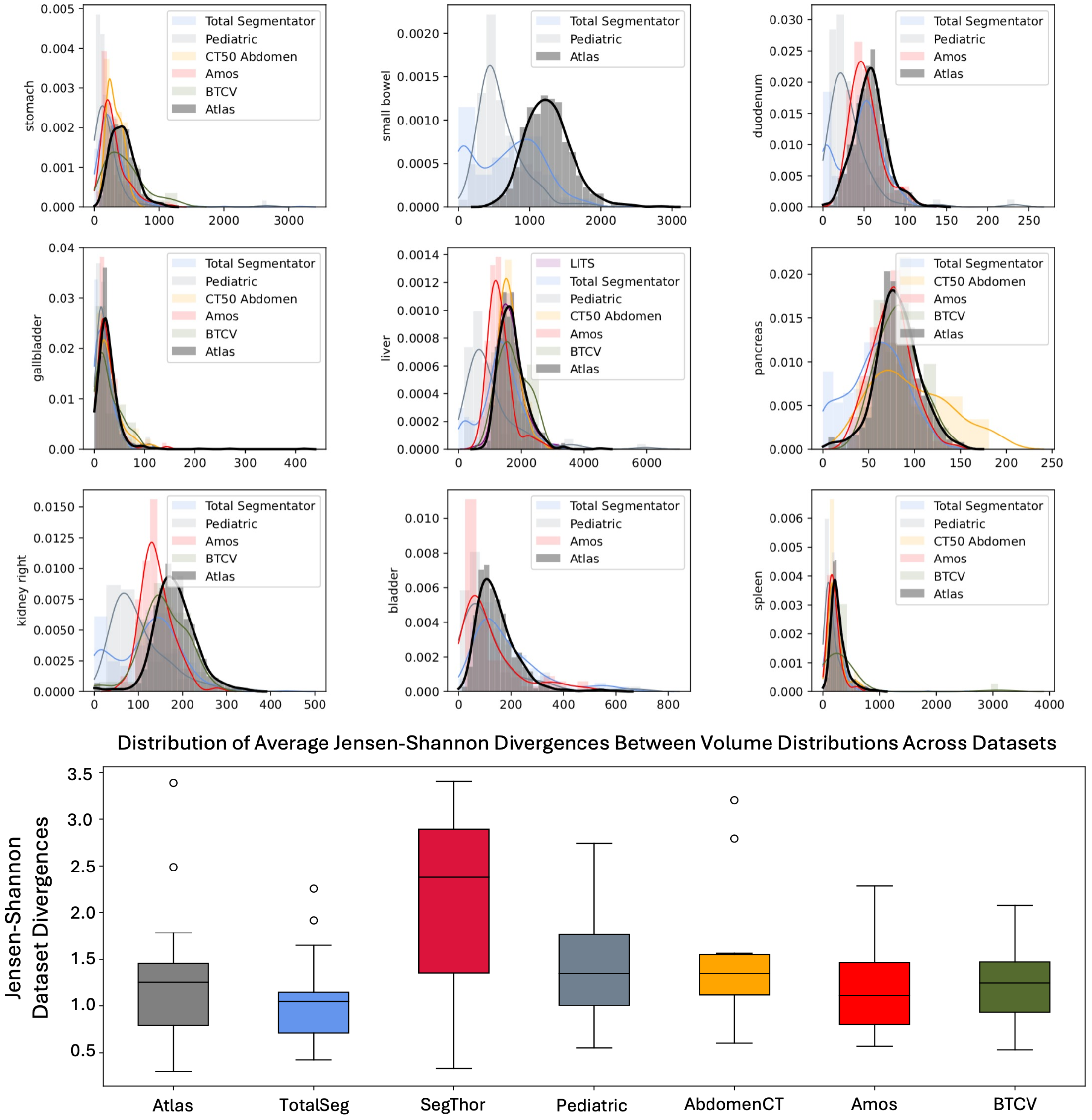


Figure 10: In the upper part, we show a comparison of the distribution of volumes in milliliters for exemplary anatomical structures. Each color indicates organ volume masks coming from different datasets. In black, we show the distributions derived from the DAP Atlas masks. To compare how well the distributions align with each other, we calculate the Jensen–Shannon divergence between each distribution and the corresponding distributions from all other datasets, and then average these divergences. This yields a mean Jensen–Shannon divergence per dataset and organ, which reflects how closely the volume statistics of a given dataset align with those of the others. Repeating this procedure across organs and datasets, we summarize the results in a box plot, thereby providing a comparative view of anatomical consistency between datasets.

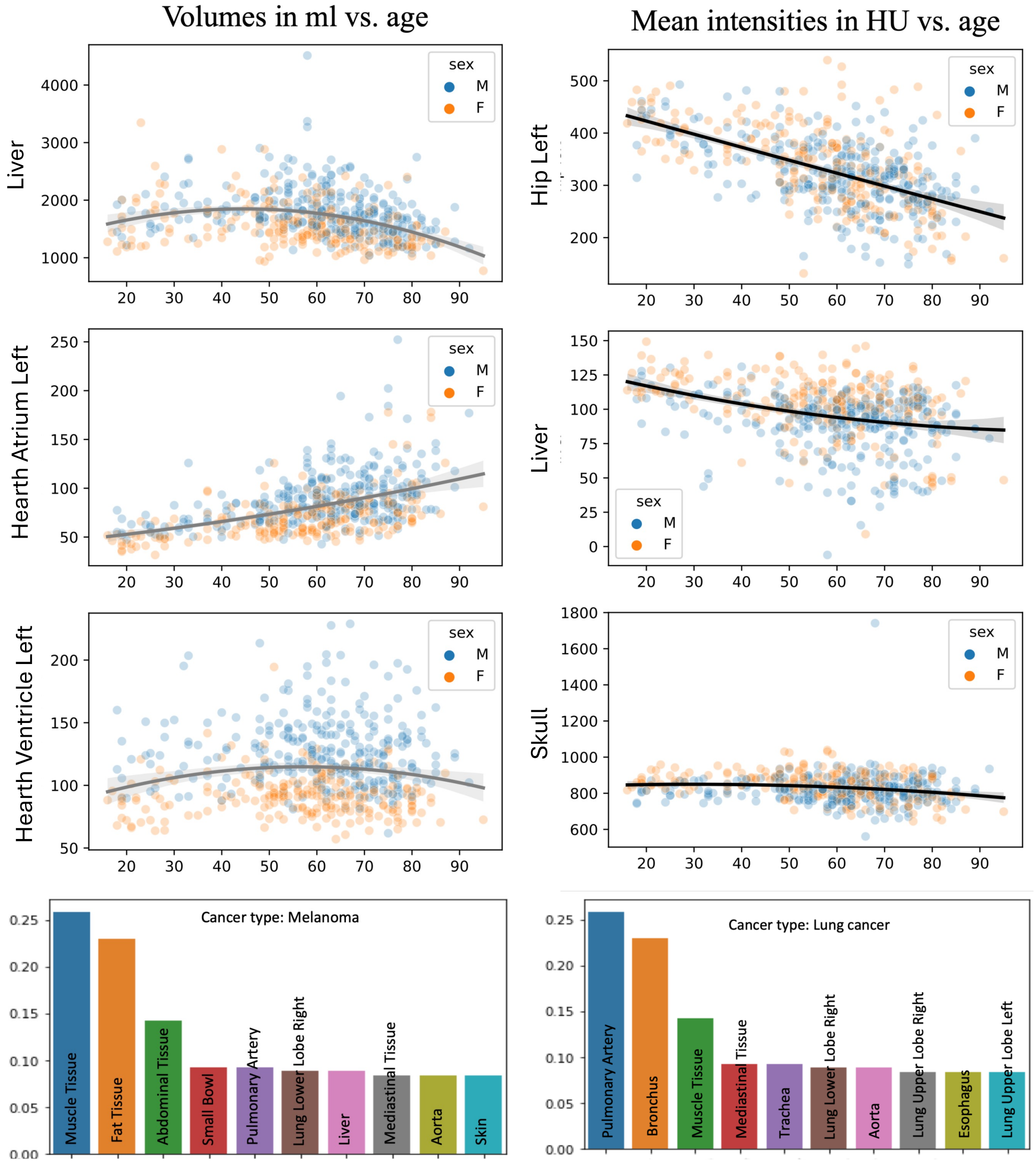


Figure 11: Age-related trends in organ volumes (scatter plots left) and mean CT intensities in HU (scatter plots right), stratified by sex. The bottom row illustrates a potential application of the Atlas dataset, showing the distribution of cancer-affected anatomical structures.

3.2 GOOD ENOUGH? AN INVESTIGATION ON THE RELEVANCE OF LABEL QUALITY

The following section is based on our work [127] (currently in submission).

We previously introduced the DAP Atlas dataset as an approach in which we automatically generate a dataset from scattered anatomical knowledge. While the overall dataset quality was rated as impressive by a radiologist, as discussed in Section 3.1.6, errors cannot be avoided for every mask in every scan. This finding is not confined to the dataset we provide, but also applies to other, more recently introduced large-scale datasets [256, 333], which contain masks of lower quality, as recent works have pointed out [20, 177].

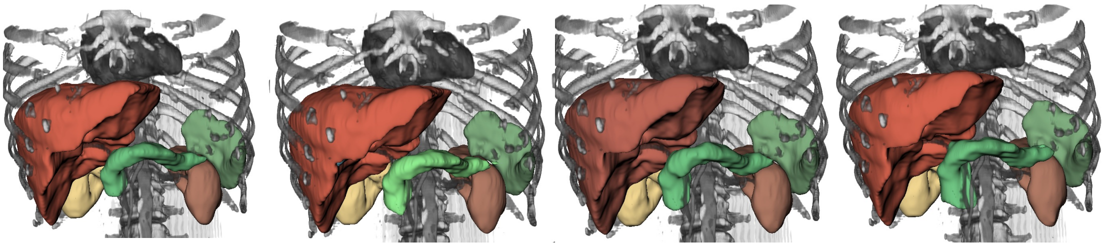


Figure 12: Can you spot the difference between these organ labels, and do you think they matter in a dataset containing tens of thousands of volumetric masks?

3.2.1 *Motivation*

12500 hours—that's how long a radiologist would need to annotate all masks in our DAP Atlas dataset [130] as derived in Section 3.1, assuming an optimistic 10 minutes per mask. Other large-scale datasets [178, 206, 256, 333] are no exception. Being clearly infeasible to annotate by hand, these works employ different strategies to ensure that the masks are of good quality: DAP-Atlas relies on algorithmic checks with radiologists for validation as explored in depth in Section 3.1.6, TotalSegmentator [333] employs a human-in-the-loop refinement, and AbdomenAtlas [178, 256] uses uncertainty-based guidance to identify automatically generated masks in need of corrections. The common approach of refining automated masks significantly reduces workload but still requires substantial radiologist involvement, with dataset curation taking weeks [256]. This raises the question: **When is improving label quality actually worth the effort?** To answer this, we analyze the impact of label quality for medical segmentation across various datasets, identifying when improvements matter and when they have minimal effect.

We create multiple versions of the same dataset of varying quality by treating the dataset as a prediction task for a wide range of models, including nnU-Net [120] and recent medical foundation models [118, 204, 333]. Comparing their predictions to the actual ground truth, we assign them a label quality and use the predicted labels to train segmentation models on them. The resulting segmentation models are evaluated across two tasks: in-domain performance and pretraining suitability, assessing how well the labels support pretraining for a different downstream task of interest. Our findings indicate label quality is crucial for in-distribution evaluation but less for pretraining with subsequent fine-tuning.

3.2.2 *Methodology*

3.2.2.1 *From Labels to Pseudo-Labels*

We first outline the creation process of pseudo-labels by defining derivatives of a base dataset as

$$D_g = \{(X^i, g(X^i))\}_{i=1}^N \tag{1}$$

where g is a neural network, capable of generating predictions $g(X^i) = \hat{Y}_g^i$ based on the input X^i. Within this work, we consider a set of pseudo-label generators $\mathcal{G}$ ={nnU-Net [120], MedSAM [204], STU-Net$_{\{\text{small, base, large, huge}\}}$ [118], TotalSegmentator [333]}, allowing us to generate a total of 7 different pseudo-label datasets (one for each $g \in \mathcal{G}$) in addition to the original dataset. To keep notations consistent, we define $\mathcal{G}^+ = \mathcal{G} \cup \{\text{base}\}$ to include the original dataset labels in this notation.

Pseudo-Label Generators:

We utilize three types of pseudo-label generators, each representing a distinct family of medical segmentation models. The first category consists of in-domain dataset generators, trained on D_{base}, thus adapting the specific annotation style of the dataset. Here, we use the nnU-Net [120] as one of the most popular and successful segmentation frameworks. We use five-fold cross-validation, training a separate model for each fold. Each model is trained on four folds and predicts the fifth. This replaces the original labels with predictions from five distinct models, forming the new dataset $D_{\text{nnU-Net}}$ composed entirely of predicted labels.

The second type of pseudo-label predictor includes non-interactive foundation models, such as TotalSegmentator [333] and the STU-Net [118] family. TotalSegmentator has established itself as a widely used segmentation model that can robustly segment large parts of the human anatomy out of the box. The STU-Net family consists of the same U-Net [268] inspired architecture across 4 different model parameter sizes trained on the TotalSegmentator dataset. The authors find that scaling model parameters tends to improve segmentation results. For our study, this poses an opportunity, as the scaled versions of the same models that are trained on the same datasets are unlikely to introduce fundamentally different types of mistakes. Instead, we expect errors to be scaled versions of those seen in smaller models, making it easier to simulate scenarios in which prediction masks are slightly improved. We explore this claim qualitatively in Figure 13.

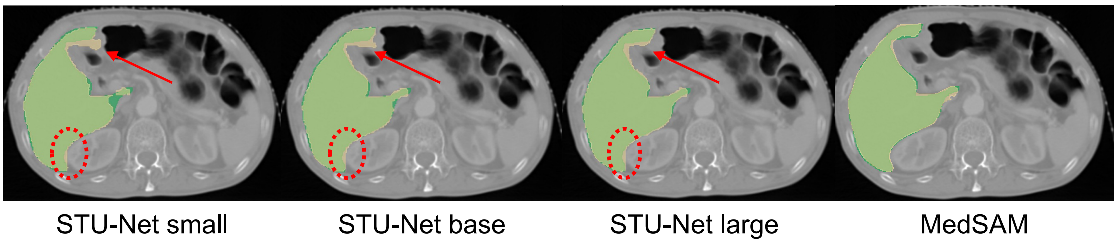


Figure 13: Overview of model prediction in yellow and the ground truth in green. STU-Net predictions tend to improve with model size, keeping the types of errors the models make constant, e.g. over-segmentation of Couinaud's liver segment VI (red circle) and iterative improvement in segment IV (red arrows). Best seen on screen with zoom.

A final remark on non-interactive pseudo-label predictors: neither the STU-Net variants nor TotalSegmentator predictions can be steered toward the desired annotation scheme. We thus remap one or multiple of the predicted labels to represent the corresponding label in D_{base}. We ignore all labels that cannot be matched. More details are outlined in Equation (3).

Finally, we utilize the interactive foundation model MedSAM [204], which provides a bridge between the in-domain dataset generators, which are perfectly able to capture the annotation style of the dataset, and the previously discussed non-interactive models, whose predictions cannot be altered. To generate predictions with MedSAM, we slice the volumetric images and generate bounding boxes around the ground-truth segmentation masks to prompt the model, allowing it to adapt to the original labels without requiring training.

Selection of base Datasets:

As datasets D, we select BTCV [168] (30 cases, 13 organs), WORD [202] (100 cases, 16 organs), AMOS [133] (200 cases, 15 organs), CT-1K [206] (1,000 cases, 4 organs), and AbdomenAtlas [178, 256] (5,200 cases, 9 organs).

These datasets are chosen based on their popularity, shared focus on abdominal organ segmentation, and increasing dataset scale. The progressive increase in dataset size and organ diversity enables a comprehensive analysis of label quality impact across multiple organs and dataset magnitudes.

3.2.2.2 *Assessing the Quality of the Pseudo-Label datasets*

For each base dataset D, we compute the different pseudo-label variants as outlined in Equation (1). Following [206], we measure label quality via Dice and Surface Dice to assess how well each pseudo-label dataset (e.g., $\text{BTCV}_{\text{nnUnet}}$) matches its original labels ($\text{BTCV}_{\text{base}}$). We report the results in Table 5 which were computed as follows: For each pseudo-label $\hat{Y}_g^i$ in D with $i \in \{1, \dots, N\}$ and each organ structure $k \in \{0, \dots, K\}$ in D, let $d_{k,i} = \text{Dice}(Y_{base}^i(k), \hat{Y}_g^i(k))$ and $s_{k,i} = \text{SurfDice}(Y_{base}^i(k), \hat{Y}_g^i(k))$ be the binary

Table 5: Overview of the quality of different pseudo-label dataset variants

		BTCV	Word	Amos	CT1k	Abd. Atlas
nnUnet	Dice	82.8±15.3	83.7±12.8	89.2±12.6	95.2±4.3	90.6±13.7
	Surf. Dice	77.9±8.8	76.5±8.7	85.2±9.2	88.9±9.4	84.4±9.2
MedSAM	Dice	63.1±20.8	56.8±18.6	59.9±23.7	72.9±13.6	39.4±20.5
	Surf. Dice	55.9±7.5	49.6±7.0	51.9±7.2	57.8±7.6	24.8±5.0
TotalS	Dice	84.2±12.0	79.3±10.3	82.9±13.7	91.9±5.5	86.6±16.4
	Surf. Dice	78.7±8.9	65.4±8.1	71.6±8.5	80.3±9.0	76.1±8.7
STU S	Dice	81.4±14.3	74.8±13.6	80.9±15.5	91.5±6.5	83.2±19.9
	Surf. Dice	73.7±8.6	57.6±7.6	66.8±8.2	78.6±8.9	70.0±8.4
STU B	Dice	83.1±12.9	75.9±12.5	82.1±15.5	92.0±6.4	84.7±18.3
	Surf. Dice	76.4±8.7	59.5±7.7	66.7±8.3	80.1±8.9	72.7±8.5
STU L	Dice	84.4±11.8	76.5±12.9	82.7±15.1	92.2±6.0	85.2±18.2
	Surf. Dice	78.1±8.8	60.0±7.7	69.6±8.3	80.6±9.0	73.8±8.6
STU H	Dice	84.4±12.3	77.2±12.1	83.0±14.6	92.2±5.8	85.8±17.4
	Surf. Dice	78.4±8.9	60.9±7.8	69.7±8.3	80.8±9.0	74.4±8.6

Dice and Surface Dice scores of organ k, where $Y^i_{base}(k) = \mathbb{1}_{Y^i=k}$ is the binary mask of the k-th anatomy label. $\hat{Y}^i_g(k) = \mathbb{1}_{\hat{Y}^i_g=k}$ is defined analogously. We first compute the organ-level mean and standard deviation by averaging over the N subjects in D separately per organ:

$$d_k = \frac{1}{N}\sum_{i=1}^{N} d_{k,i}, \quad \sigma_{d,k} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\, d_{k,i} - d_k \,)^2}, \tag{2}$$

and analogously $s_k, \sigma_{s,k}$ for Surface Dice.

To aggregate the organ-wise scores we average d_k across organs. To ensure a fair comparison, we only consider anatomical structures that all generators g are capable of predicting and ignore the others. This is necessary since the non-interactive pseudo-label predictors only offer a fixed set of labels that cannot be influenced. These are, however, only minimal adjustments affecting the class *Prostate* in Amos and *Rectum* in Word, which are excluded from the calculation of the average. Formally, we define the set of labeled anatomical structures in a dataset D as $\mathcal{L}_D = \{l_0, l_1, \dots, l_{K_D}\}$, where l_0 typically represents the background. The relevant set of labels used for pseudo-labeling is denoted as $\tilde{\mathcal{L}}_D \subseteq \mathcal{L}_D$. To derive this set, we first remove l_0 since we ignore the background prediction. Then, we consider only labels that each pseudo-label generator $g \in \mathcal{G}$ is capable of predicting, resulting in the set of common labels:

$$\tilde{\mathcal{L}}_D = l \in \mathcal{L}_D \setminus l_0 \mid \forall g \in \mathcal{G}, l \in \mathcal{L}_g \tag{3}$$

where $\mathcal{L}_g$ represents the set of labels that generator g can predict.

$$d = \frac{1}{|\tilde{\mathcal{L}}_D|} \sum_{k \in \tilde{\mathcal{L}}_D} d_k, \quad s = \frac{1}{|\tilde{\mathcal{L}}_D|} \sum_{k \in \tilde{\mathcal{L}}_D} s_k, \tag{4}$$

To aggregate the standard deviations σ_d and σ_s of individual organs, we compute the root-mean-square of the organ-level standard deviations over $\tilde{\mathcal{L}}_D$.

When analyzing the results, we observe various patterns: (1) There is a clear improvement pattern from STU S → B → L → H, though the gains become smaller. The largest improvements are seen between STU S and STU B, with smaller returns for L and H variants across both metrics. This quantitatively confirms the previously discussed desire to simulate slight mask improvements. (2) nnUNet outperforms on most datasets, achieving Dice scores from 82.8% to 95.2%, benefiting from direct training on each dataset's annotation style. Despite the non-interactive foundation models being trained on larger datasets, they generally lag behind. An exception is BTCV (30 images), where they surpass nnUNet, but this advantage vanishes on larger datasets, emphasizing nnUNet's superior scaling. (3) MedSAM, as a candidate for an interactive foundation model, fails to exhibit its unique advantage of adapting to the annotation style through interactions, despite benefiting from extensive pre-training. A potential reason for its consistently inferior performance is that it natively operates in 2D, which prevents it from capturing the volumetric nature of the data, resulting in artifacts when stacking its 2D predictions back into 3D. (4) TotalSegmentator and STUNet$_{\text{huge}}$ perform about equally well, providing strong performance despite not being fine-tuned to the data or adapting to dataset-specific annotation styles.

3.2.3 *Experiments and Results*

We now turn to an evaluation of the impact of dataset quality on different downstream tasks. We consider two distinct tasks: In-Domain evaluation and pre-training suitability.

Training Setup:

We chose DynUnet [121] as our model of choice, as it offers a flexible and well-proven architecture serving as the workhorse of the nnU-Net [120] framework. Let $f_{D_g}, g \in \mathcal{G}^+$ be a deterministically trained DynUnet on the Dataset version of D that has been generated by g. To mitigate any influence on model training except for the data, we use a completely deterministic training for 1000 epochs where each epoch consists of exactly 250 mini-batches, thereby oversampling the dataset if needed. For each dataset, we resample images to their respective datasets' median resolution before extracting patches. The patch sizes are specific for each dataset D, but remain constant across the different versions $D_g \forall g \in \mathcal{G}$. The network weights are updated using the AdamW [199] optimizer over a cosine-annealing learning rate scheduler starting from a base learning rate of $1e-3$. We leverage random 80/20 splits, setting aside 80% of the data for training and 20% for evaluation. The splits are consistent across all versions of D.

3.2.3.1 *In-Domain Evaluation*

For the task of in-domain evaluation, we consider the setup, where a model $f_{D_g}, g \in \mathcal{G}$ is evaluated against the dataset D_{base}. This represents a scenario where a model is only trained on pseudo-labels but has to perform inference on the original labels. This is a setting commonly found where given target structures should be segmented; however, training labels are not available, and researchers may consider using pseudo-labels.

We report the results in Figure 14 and Figure 15 and summarize our findings as follows. (1) Overall, we see a strong correlation between label quality and network performance. (2) Models trained on nnU-Net labels consistently outperform others, achieving superior Dice and SDice scores across all datasets, highlighting the benefits of training on in-domain data, and thereby adapting possible annotation schemes. (3)$f_{D_{MedSAM}}$ uniformly underperforms, with notably poor performance on Abdomen Atlas. (4)Among the STU variants and TotalSegmentator, it is dataset-dependent whether we see an increase in performance with an increase in label quality. Although a promising, albeit non-monotonic, pattern can be observed in the Word dataset, on other datasets, such as Amos or Abdomen Atlas, small label improvements are not necessarily translated into better performance. We hypothesize that one reason could be the type of noise that is generated by the pseudolabel generator. While random noise can be learned to be ignored by a model, systemic noise leads the model to adapt the systemic error. (5) With increasing dataset size, we observe that marginal performance improvements from enhanced dataset quality exhibit diminishing returns. This trend is evident in the changing slope of the regressed dataset-quality model performance relationship shown in Figure 15. We hypothesize that with a sufficiently large dataset, models become capable of extracting the underlying signal while filtering out inconsistent noise, thereby rendering smaller annotation variations negligible. Beyond a certain threshold, we expect pseudolabels to serve as viable approximations of the true ground truth, of which human-annotated labels are themselves high-quality approximations, enabling models to achieve comparable performance whether trained on pseudolabels or original training data. However, this relationship likely varies significantly across medical imaging domains.

Limitations of this study:

While our findings may hold true for anatomy segmentation tasks where structural boundaries are relatively well-defined and consistent, pathology presents a fundamen-

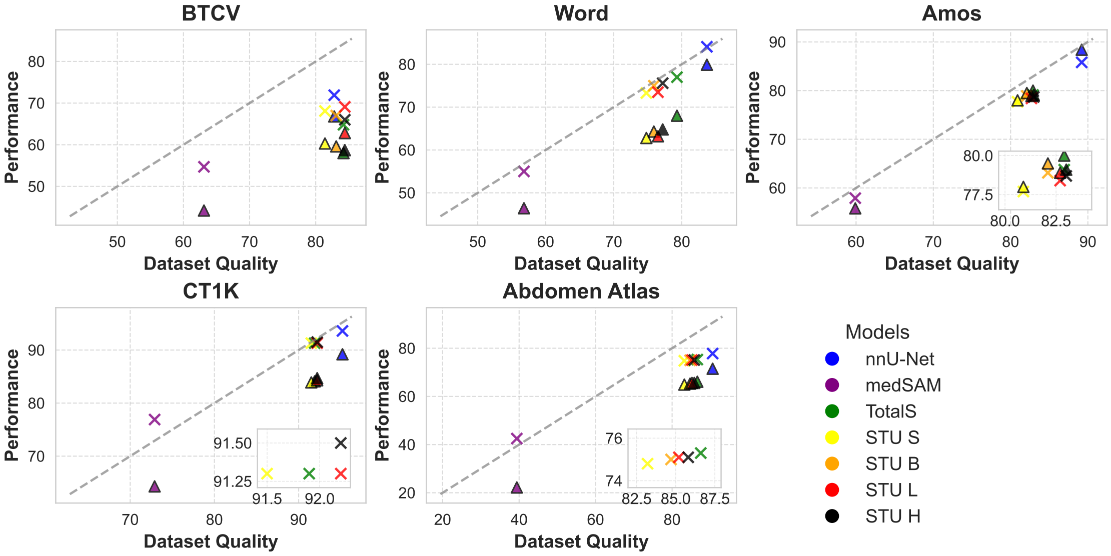


Figure 14: In-Domain evaluation results for Dice ("×") and Surface Dice ("Δ"). We include dashed $y = x$ for reference and zoom into areas if markers are cluttered.

tally different challenge. In pathological imaging, subtle morphological changes often carry critical diagnostic significance; yet, these nuanced features may not be adequately captured by global metrics, such as the Dice or Surface Dice coefficient, when aggregated across an entire dataset. Consequently, the tolerance for label noise and the viability of pseudolabel approximations may be substantially lower in pathological contexts, where precision at the pixel level can significantly impact diagnostic accuracy and clinical outcomes.

3.2.4 *Pre-training Suitability*

Pre-training a model on a dataset before fine-tuning it on a different dataset is a second use case we investigate. We utilize the previously trained networks $f_{D_g}, g \in \mathcal{G}^+$, which we will fine-tune on two downstream tasks: segmenting thoracic organs in SegTHOR [167] and segmenting abdominal organs in the Flare Dataset [207]. We employ the same training setup as previously and keep the crop size, target resolution, and normalization schemes constant with respect to the pre-trained model. We present the results for SegTHOR in Table 6 and for Flare in Table 7. Cells, where training did not converge, are marked with "—". We compute the quintiles for the data separately by column and metric, and color the cells according to the quintile they fall into, from worst (red) to best (green). In both tables, the columns indicate which datasets D were used, and the indices of the rows indicate the pseudo-label generators g on whose

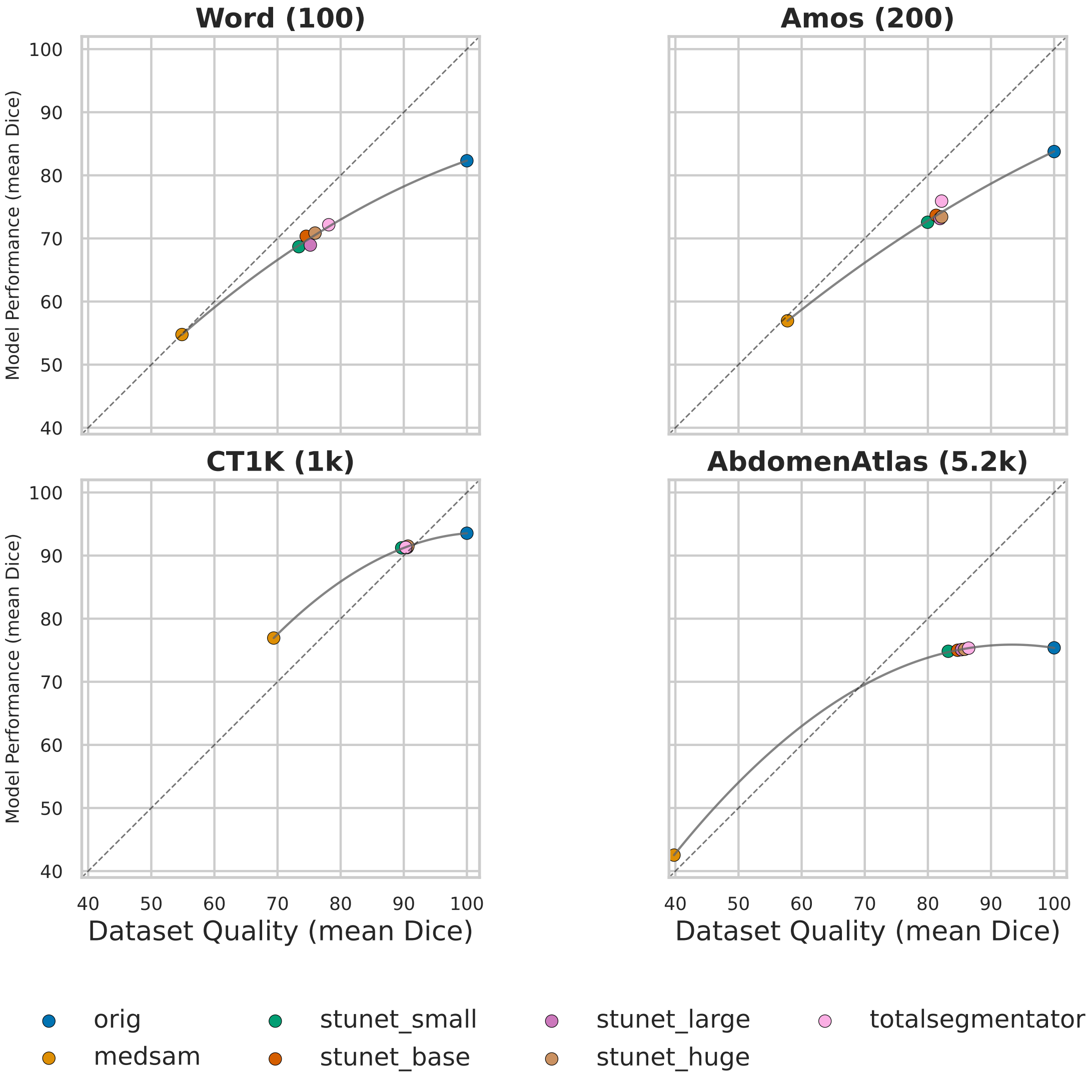


Figure 15: In-domain evaluation of pseudo-labeled training on original labels. For each dataset (2 × 2 grid), models f_{D_g} trained solely on pseudo-labels from generators $g \in \mathcal{G}$ are evaluated against the base dataset D_{base} with original annotations. Points denote generators (consistent colors across panels). Axes are equal and shared across panels to enable direct comparison. A quadratic regression (solid) summarizes the relationship between dataset quality x and model performance y. The dashed line indicates the identity $y = x$. With increasing dataset size, we observe, marginal gains from higher pseudo-label quality diminish, as reflected by the flattening slopes in the dataset-quality vs. performance regressions

Table 6: Finetuning Evaluation results of the SegTHOR dataset.

		BTCV	Word	Amos	CT1K	Abd.Atlas
no pretrain	Dice	87.4 ± 2.0	87.0 ± 2.3	88.2 ± 2.2	86.5 ± 3.0	83.5 ± 8.6
	SDice	77.5 ± 2.8	79.7 ± 3.2	88.5 ± 2.7	74.2 ± 3.0	62.6 ± 13.4
base	Dice	87.6 ± 2.0	88.8 ± 2.0	89.7 ± 1.6	87.1 ± 2.7	87.3 ± 5.0
	SDice	78.5 ± 2.7	82.8 ± 2.7	90.2 ± 2.2	75.0 ± 2.7	70.5 ± 7.6
nnU-Net	Dice	87.7 ± 1.9	88.1 ± 2.1	89.2 ± 1.8	88.0 ± 2.8	87.2 ± 5.8
	SDice	79.8 ± 2.6	81.8 ± 2.7	90.1 ± 2.2	76.3 ± 2.8	70.9 ± 9.1
MedSAM	Dice	87.1 ± 2.2	87.6 ± 2.3	89.2 ± 1.9	88.0 ± 2.8	87.0 ± 5.0
	SDice	79.7 ± 2.8	80.8 ± 3.0	89.7 ± 2.4	76.1 ± 2.8	69.8 ± 8.2
TotalS	Dice	87.6 ± 2.0	88.5 ± 2.1	89.1 ± 1.9	87.5 ± 2.7	87.6 ± 4.7
	SDice	79.4 ± 2.6	82.7 ± 2.7	89.6 ± 2.4	76.0 ± 2.7	71.7 ± 7.5
STU S	Dice	87.5 ± 2.0	89.0 ± 1.8	89.3 ± 1.7	87.6 ± 2.8	87.2 ± 4.8
	SDice	79.6 ± 2.8	83.0 ± 2.7	89.5 ± 2.2	76.4 ± 2.8	70.2 ± 8.8
STU B	Dice	87.9 ± 1.8	88.0 ± 2.2	89.4 ± 1.7	–	87.0 ± 6.1
	SDice	79.6 ± 2.7	81.4 ± 2.8	90.1 ± 2.3	–	70.6 ± 9.6
STU L	Dice	87.3 ± 2.0	88.7 ± 2.0	89.0 ± 1.9	87.4 ± 2.9	87.5 ± 5.3
	SDice	79.3 ± 2.6	82.6 ± 2.6	89.7 ± 2.3	76.1 ± 2.9	70.9 ± 7.7
STU H	Dice	87.5 ± 2.0	88.0 ± 2.1	89.2 ± 1.8	88.0 ± 2.6	87.3 ± 5.0
	SDice	79.6 ± 2.8	81.5 ± 2.6	89.7 ± 2.3	76.8 ± 2.6	69.7 ± 8.7

pseudo-labels the model $f_{D_g}, g \in \mathcal{G}$ was pre-trained. Finetuning is always performed on the original labels.

We summarize our findings as follows: (1) Pre-training consistently improves over initializing models from scratch: Dice and Surface Dice performances are improved, and their variances are reduced. (2) The data quality still matters for pre-training but to an almost negligible extent. While nnUNet labels and the original labels yield the best results on average, the performance gap to MedSAM, previously an outlier with the worst performance, has narrowed. MedSAM remains the weakest, but it is now well within the range of other pre-training approaches. This indicates that models transfer rather general concepts from the pre-training dataset compared to fine-grained information. (3) Optimizing dataset processing is more crucial than pre-training. This can be seen as scores vary more by column than by row.

One limitation of the previously reported results is the lack of comparability across datasets. Since we utilize models developed during in-domain experiments, as described in Section 3.2.3.1, we adjust the data processing of the downstream task to align with the model configurations of the in-domain models. This approach constrains our

Table 7: Finetuning Evaluation results of the Flare dataset.

		BTCV	Word	Amos	CT1K	Abd.Atlas
no pretrain	Dice	93.4 ± 2.0	93.5 ± 1.9	93.3 ± 2.1	93.5 ± 2.1	93.2 ± 2.1
	SDice	90.5 ± 2.8	91.5 ± 2.6	93.8 ± 2.5	89.4 ± 2.8	86.9 ± 3.0
base	Dice	93.8 ± 1.9	93.8 ± 1.8	93.5 ± 1.9	94.0 ± 1.8	93.8 ± 1.7
	SDice	91.2 ± 2.7	92.0 ± 2.5	94.0 ± 2.4	90.0 ± 2.8	87.7 ± 2.8
nnU-Net	Dice	93.7 ± 1.9	94.0 ± 1.8	93.7 ± 1.8	94.0 ± 1.8	93.9 ± 1.8
	SDice	91.0 ± 2.7	92.2 ± 2.4	94.4 ± 2.2	90.2 ± 2.7	87.8 ± 2.8
MedSAM	Dice	93.5 ± 2.0	93.8 ± 1.8	93.5 ± 2.0	93.7 ± 2.0	93.6 ± 1.8
	SDice	90.8 ± 2.7	92.0 ± 2.5	94.0 ± 2.5	89.8 ± 2.8	87.4 ± 2.9
TotalS	Dice	93.8 ± 1.8	94.0 ± 1.8	93.7 ± 1.9	94.0 ± 1.8	93.7 ± 1.8
	SDice	91.2 ± 2.5	92.2 ± 2.5	94.4 ± 2.3	90.0 ± 2.6	87.3 ± 2.9
STU S	Dice	93.7 ± 1.8	93.9 ± 1.8	93.6 ± 1.9	93.9 ± 1.8	93.7 ± 1.9
	SDice	91.0 ± 2.6	92.2 ± 2.4	94.2 ± 2.3	90.1 ± 2.6	87.5 ± 2.9
STU B	Dice	93.8 ± 1.8	94.0 ± 1.7	–	–	93.7 ± 1.8
	SDice	91.1 ± 2.6	92.2 ± 2.4	–	–	87.7 ± 2.8
STU L	Dice	93.7 ± 1.9	93.9 ± 1.8	93.8 ± 1.8	94.0 ± 1.8	93.8 ± 1.8
	SDice	91.1 ± 2.7	92.1 ± 2.5	94.4 ± 2.2	90.1 ± 2.6	87.8 ± 2.9
STU H	Dice	93.7 ± 1.9	94.0 ± 1.7	93.6 ± 1.9	93.9 ± 1.9	93.8 ± 1.8
	SDice	91.1 ± 2.6	92.2 ± 2.4	94.2 ± 2.3	90.2 ± 2.6	87.6 ± 2.9

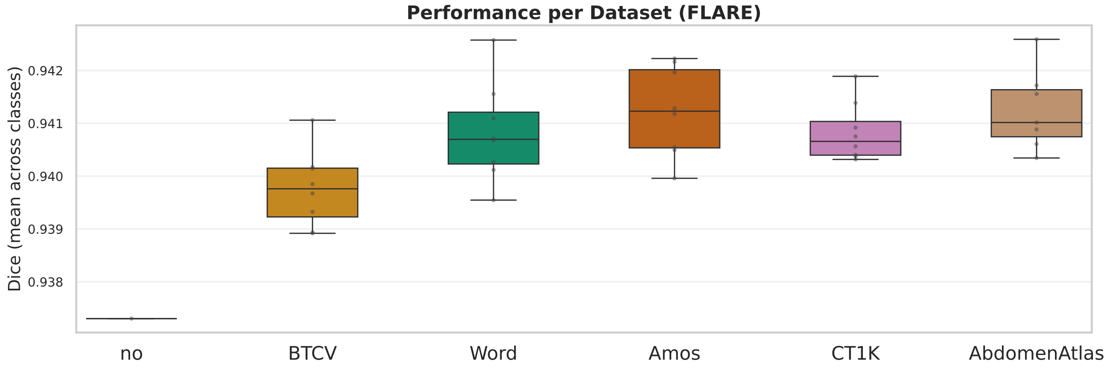


Figure 16: Boxplot comparing the effect of pretraining on different variants of base datasets across different pseudolabel generators, including original data. The models were pretrained on the respective source dataset variants as generated by the pseudolabel generators and subsequently fine-tuned on the clean Flare [207] dataset, on whose test set the respective model performance was evaluated.

ability to compare results across pretraining datasets, as model configurations can significantly impact performance [120]. To address this, we repeat the pretraining procedure using the desired target configuration of the downstream task, which we then use to prepare the pretraining datasets. For the Flare dataset, we now aggregate the results across different datasets and report the results in a boxplot in Figure 16. Interestingly, we find that larger datasets, such as CT1k [206] or AbdomenAtlas [178, 256], do not necessarily hold an advantage over the smaller dataset Amos [133], at least under our chosen fixed computational budget setting. When comparing the results across different predictors, outlined in Figure 17, we observe a similar pattern compared to what we previously discussed in in Section 3.2.4 and in Tables 6 and 7: label quality seems to play a limited role. Even models trained on the medsam-generated pseudolabels perform in the same league as those trained on much higher-quality pseudolabels. We confirm these findings for a second dataset SegThor [167] in Figures 38 and 39 in the Appendix C.2.

In summary, we find that for pre-training it is not worth improving the label quality. Quality matters only if major improvements can be achieved. Small improvements will not be translated into better performance.

3.2.5 *Discussion*

Our findings challenge prevailing assumptions about the necessity of large-scale, perfectly annotated datasets in medical segmentation. The strong correlation between label quality and in-domain performance across datasets confirms intuitive expectations, yet reveals diminishing returns as dataset size increases beyond ~1,000 cases. More surprisingly, for pretraining applications, label quality exhibits minimal impact:

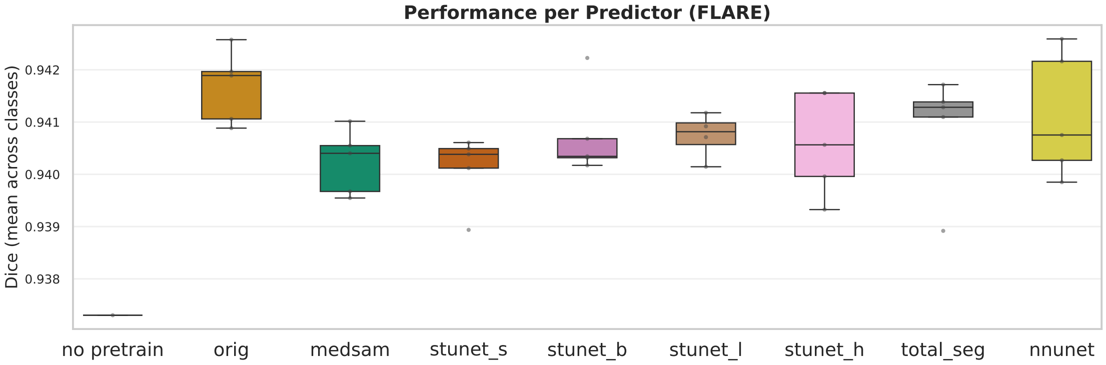


Figure 17: Boxplot comparing the effect of pretraining across datasets for different pseudolabel generators. The models were pretrained on the respective source dataset variants as generated by the pseudolabel generators and subsequently fine-tuned on the clean Flare [207] dataset. The results are aggregated by comparing the influence of the pseudolabel generator.

Even poor-quality MedSAM pseudo-labels provide comparable transfer learning benefits to high-quality annotations.

Implications for Dataset Creation: The medical AI community may be overinvesting in large-scale dataset curation. Our results suggest that moderate-sized, well-curated datasets with consistent annotation protocols may be a more cost-effective and meaningful strategy than massive collections purely for the sake of scaling. This contradicts the recent trend toward ever-larger datasets [178, 256, 333] and supports more strategic resource allocation unless clear use cases for these large datasets can be identified.

.

3.3 ADAPTING AND IMPROVING LABELS USING THE LIMIS ARCHITECTURE

Models trained on imperfect labels—which we now know are prevalent in large-scale medical datasets [256, 333] may make mistakes, but even models trained under perfect labels can make mistakes due to annotation shifts (e.g., liver vessels are treated as part of the liver instead of a separate class), generalization issues, and domain-specific requirements. When segmentation results do not meet clinical expectations, how can we efficiently adapt them?

The following section is based on our ISBI 2025 work [105] and introduces LIMIS: A **L**anguage-based **I**nteractive **M**edical **I**mage **S**egmentation framework, which allows users to generate and, more importantly, refine segmentation masks using natural language, providing a hands-free solution for scenarios where traditional physical input device-based interactions are impractical.

Model	Interactive	Physical Interact.	Lang.-based Seg.	Lang.-based Interact.
nnUNet [120]	✗	-	-	-
TotalSeg. [333]	✗	-	-	-
SAM-based [154]	✓	✓	✗	✗
ScribblePromt [338]	✓	✓	✗	✗
GroundedSAM [264]	✓	✗	✓	✗
LIMIS (ours)	✓	✗	✓	✓

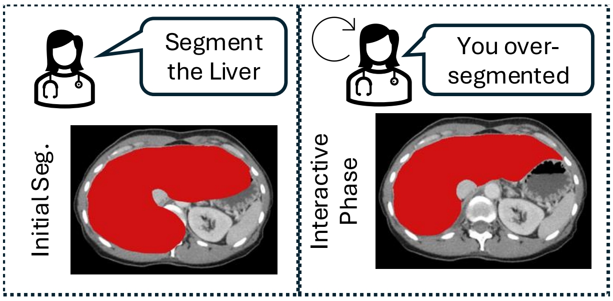


Figure 18: Comparison of LIMIS with prior work. LIMIS introduces a unique, purely natural-language-based segmentation and interaction strategy, extending beyond conventional interactive or modality-specific approaches. On the right, we show an exemplary workflow.

3.3.1 *Motivation behind LIMIS*

Interactions in the medical field are currently limited to direct physical interactions between a physician and a model, such as scribbles [338] or clicks [46] that are typically performed using mouse movements or mouse clicks. A downside of this approach is that these methods cannot be used in situations where physicians need to use their hands to perform treatments or surgeries while depending on precise, problem-tailored segmentations. Typical examples in the clinical routine are orthopedic surgeries such as the insertion of implants [165] that require intraoperative CT

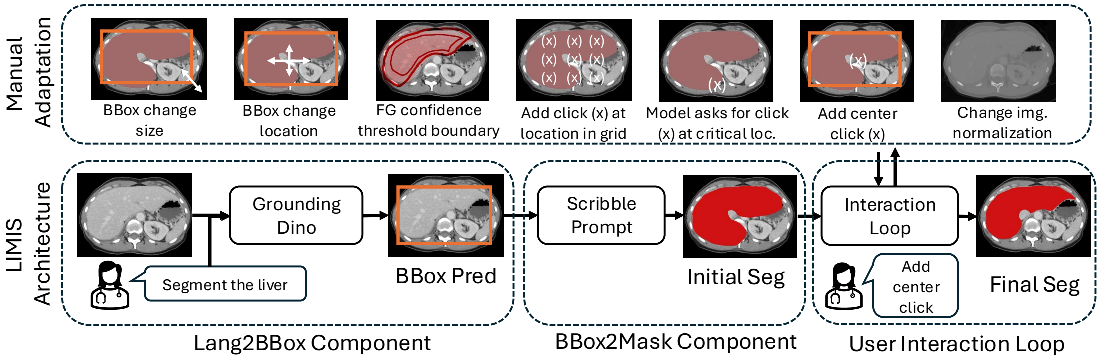


Figure 19: Top: Manual Language-based Adaptation options. Bottom: LIMIS flowchart showing user input processing from language prompt to final mask via Grounding DINO (Lang2BBox), ScribblePrompt (BBox2Mask), and User Interaction Loop.

images, real-time imaging in endoscopy [65, p. 443–450] or real-time X-rays during cardiac catheterization [215]. To address the shortcomings of current physical interactive segmentation models, we pioneer the development of a model that can work with natural language. We address the primary challenge of designing a system that effectively utilizes natural language for segmentation and interaction tasks. In the following, we describe our efforts to make significant progress towards this goal by first developing a framework that works with text-based inputs, laying the groundwork for future adaptation to spoken language, which, given the robust capabilities of existing Voice2Text models, can expected to be seamless.

3.3.2 *Methodology*

LIMIS consists of three major components: the Language to Bounding Box component (Lang2BBox), which works with the Bounding Box to Segmentation component (BBox2Mask) to generate an initial segmentation, and the User Interaction Loop. Figure 19 shows the structure of the LIMIS architecture, including some of the manual user interactions.

3.3.2.1 *Generating an Initial Segmentation from Language*

To generate an initial segmentation mask from language input, we draw inspiration from the Grounded SAM [264] architecture, which has already been explored for colonoscopy [29] or X-ray [258]. Contrary to these works, we do not keep the standard Grounded SAM architecture but adapt both its components: SAM [154] since it has been shown to perform poorly on non-optical medical images such as radiographic images [338], and Grounding DINO [191].

To obtain an initial segmentation, we first generate a bounding box from a language prompt in the Lang2BBox component. To achieve this, we adapt the text-based object detector Grounding DINO [191] to the medical domain using the parameter-efficient fine-tuning method LoRA [113]. This LIMIS component predicts a bounding box around the target object. In the BBox2Mask component, we use the predicted bounding box as a prompt to the ScribblePrompt [338] model, which is a medical adaptation of SAM [154], to predict an initial segmentation mask.

3.3.2.2 *Segmentation Refinement through User Interactions*

The third component of LIMIS is the User Interaction Loop, allowing refinements of the initial segmentation mask via user interactions. It starts by applying a default adaptation to the image and segmentation mask. Users then assess if this improves the segmentation mask and choose whether to keep it. This default strategy normalizes the CT image based on the target organ's typical radiological visualization parameters, e.g., using liver-specific CT window settings. The strategy further expands the bounding box by 10 pixels on each side. The choices for these default values are ablated in Section 3.3.3.2.

After applying the default options, users have two methods to address potential segmentation mask errors:

- **Manual Adaptation**: Adjust the segmentation mask through manual interactions.
- **Automated Multi-Step Strategies**: Choose from four predefined automated strategies designed to correct common segmentation issues.

Throughout the segmentation process, users can decide after each interaction whether to continue with the updated mask or revert to any previous version. The final segmentation mask can be selected from any step, and it does not need to be the one generated in the last interaction.

Manual Adaptation via Interactions: LIMIS offers manual interactions inspired by physical click-based interactions and active learning regimes:

- **Bounding Box Changes**: Shift location or change size.
- **Confidence Threshold**: Change the threshold determining if critical pixels are part of the foreground mask.
- **Click in Grid**: Add a foreground/background click in one of 16 locations organized as a regular grid.
- **Critical Region Decision**: The system asks the users to decide for specific critical points if these belong to the foreground structure or the background.

- **Center Click**: Add a foreground click in the center of the bounding box.
- **Change Normalization**: Choose a new CT visualiztion window (location & width) for image normalization.
- **Generate Examples**: Show exemplary interactions.
- **Remove Component**: Remove a connected component.
- **Ensemble**: Combine the segmentation masks of the following interactions: box size change, center click, and change of normalization.

Problem-oriented, guided multi-step Interactions:
Besides the manual interactions, four problem-oriented, predefined multi-step adaptations guide the user. We show an example for this in Figure 18 on how to refine the initial segmentation mask:

- **Wrong image part segmented**: Add center click, adjust normalization, and add grid points.
- **Target oversegmented**: Increase the foreground confidence threshold and add critical points and grid points.
- **Target Undersegmented**: Increase BBox, reduce foreground confidence threshold, and add critical points.
- **Target has low HU-values**: Adapt image normalization.

In each of the four suggestions, the predefined manual interactions guide the users, thereby streamlining the segmentation process and helping the users to familiarize themselves with the effects of the manual interactions used during the automated processes.

3.3.2.3 *Adaptation Strategy Grounding DINO (Lang2BBox)*

Within the following subsection, we outline our proposed adaptation strategy of the non-medical Grounding DINO object detector to the medical domain.
Changes to Network Structure: We use the SOTA parameter-efficient fine-tuning approach LoRA [113] to adapt Grounding DINO to the medical domain. Compared to other domain adaptation methods, such as adapters, it does not add any additional inference time. We include LoRA layers in the self-attention and deformable self-attention layers within the Grounding DINO architecture.
Data: We use three publicly available medical CT datasets for this work: Our self-developed DAP Atlas [130], TotalSegmentator [333], and WORD [202]. In this work, we only use the anatomical structures available in all three datasets: esophagus, stomach, duodenum, colon, gallbladder, liver, pancreas, kidney left, kidney right, bladder,

and spleen. The pooled dataset is split into 80% for training, 10% for validation, and 10% for testing, ensuring no overlap in validation and test set with images seen by ScribblePrompt during its training. Each dataset is initially split into 80% training, 10% validation, and 10% testing. The resulting subsets are then pooled across all datasets, maintaining the same proportions. We make sure our test and validation sets have no overlap with images used by the authors during the ScribblePrompt model training

Data Pre-Processing: Images are pre-processed by slicing CT volumes into 2D images along the transverse plane. We clip the HU-values to the 0.5 and 99.5 percentiles. We normalize using the mean and standard deviation of the foreground pixels. To address dataset differences, we commit to a common pixel spacing, image size, and image orientation. As data augmentations, we use image translations, rotations, and scaling with an individual probability of 10%. The range of rotation is -10.3° to 10.3°, the translation up to 10 pixel,s and the scaling factor is between 0.9 and 1.1.

Language Prompt Generation: The training of Grounding DINO requires a language input, which we model as a sequence of label names that consists of two parts. The first part is the label names of the organs present in the image. We further add random label names from all training classes that are not present in the image, simulating noise in the language prompts. All label names in the prompt are shuffled randomly.

Loss Function and Hyperparameters. The loss function and most hyperparameters are chosen according to [191]. A detailed summary of the ablated training configuration is shown in Table 8.

3.3.3 *Experiments and Results*

3.3.3.1 *Grounding DINO: Implementation & Evaluation*

The fine-tuning of Grounding DINO was conducted on three NVIDIA RTX 6000 GPUs with an individual batch size of 64 per GPU, yielding a total batch size of 192. The model achieved a mean Average Precision (mAP) of 0.54, with mAP@50 at 0.80 and mAP@75 at 0.58.

Ablations We ablated the usage of augmentations (augm), the learning rate (lr), and the number of additional label names that were added to the text prompt (num add lab). Table 8 shows the influence of these hyperparameters on the results of the training. Configuration 1 achieves the highest mAP. We find that applying augmentations generally leads to improved results, and using a greater number of random label names outperforms using fewer.

3.3.3.2 *ScribblePrompt: Implementation & Evaluation*

We evaluate ScribblePromt [337] as our BBox2Seg component for different configurations. We compare feeding the entire image with its bounding box to the model as well as the image cropped to the bounding box plus a small margin around it with the

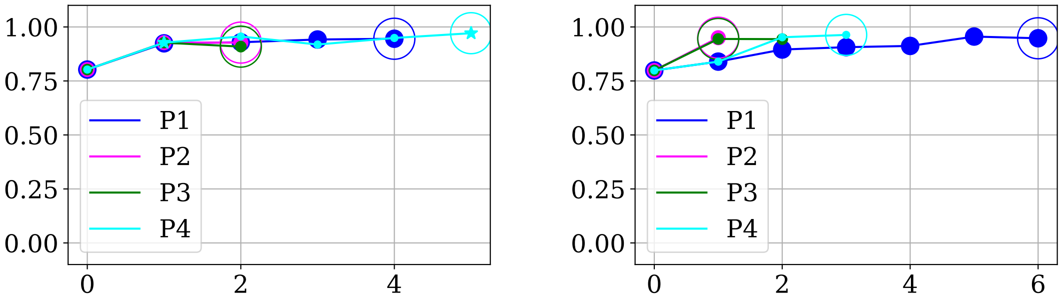
Table 8: Tested hyperparameter configurations for Grounding DINO on the validation dataset.

Config	augm	lr	num add lab	mAP
1	yes	1e-4	8	**0.541**
2	yes	1e-4	2	0.540
3	no	1e-4	8	0.525
4	no	1e-4	2	0.510
5	yes	1e-5	2	0.499

latter setting leading to significantly higher Dice scores (53% vs. 58%) across all segmented organs. We further identify that using common radiologist CT visualization windows as the input to ScribblePrompt boosts performance from 58% Dice to 63%. Finally, we investigate if the predicted bounding box should be enlarged by default by a small number of pixels. We find that on average increasing the bounding box by 10 pixels on each side improves the performance to 66% Dice. Enlarging the box further to 20 pixels per side decreases the performance significantly to 54% Dice indicating a worse localization cue by the enlarged bounding box. We show the effect of the stated default option qualitatively in Figure 21 (default).

3.3.3.3 *Interaction Loop: Evaluation via User Study*

The third component of the LIMIS architecture is the User Interaction Loop. We evaluate its performance via a user study with four participants: Two radiologists, one medical doctor, and one medical student. The users had 10 minutes to familiarize themselves with the system and were then asked to segment as many images as possible with the best possible result in 50 minutes. The users should use a maximum of 5 minutes per image and move on to the next image if two consecutive interactions did not improve the segmentation mask. We present the users with a series of CT images from our test set in which they are tasked to segment one anatomical structure. We design the user interaction interface as a GUI facilitating using the system for non-technical users. During the user study, the participants collectively annotated 63 images. We evaluate the results of the user study and find that for 41 images (65%), the final segmentation had a higher Dice score than the initial segmentation. The average Dice improvement for these images was $(6 \pm 5.13)\%$. Around 21% of the images had a lower final Dice score $(-2 \pm 2)\%$ and 14% of the images resulted in identical Dice scores pre- and post-interactions. Overall, the Dice score change was $(4 \pm 7.0)\%$. It however has to be pointed out that the participants were not forced to submit the mask after the last interaction but were allowed to submit any intermediate and even the initial prediction. Thus, some Dice score drops may reflect differing expert opinions, not system weaknesses. Additionally, it has to be acknowledged that the overall



(a) Atlas, bladder

(b) TotalSegmentator, liver

Figure 20: Dice score over interaction steps for two images. Step 0 is the initial mask; if "default" was accepted, it's step 1. Big circles mark the user's final chosen mask. Stars indicate when a non-latest step was adapted, marking both the adapted and resulting steps.

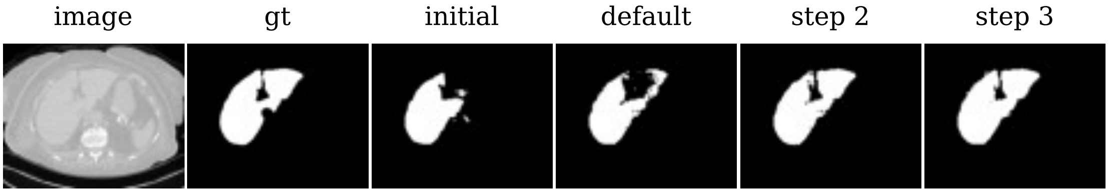


Figure 21: Liver segmentation mask over iteration steps. The first image shows the CT scan, the second the ground truth (gt), and the third the initial LIMIS prediction. "Default" presents the mask after the default option, and the last two images show masks from steps 2 and 3.

performance of LIMIS is limited by the ScribblePrompt foundation model used as the BBox2Mask component.

In Figure 20 we show the Dice scores change over the iteration steps when tasked to segment the bladder (left) within a sample taken from the DAP Atlas [130] dataset and the liver (right) from a TotalSegmentator [333] sample. A qualitative example of the change of the segmentation mask is shown in Figure 21.

We evaluate the usability of LIMIS with the NASA TLX and the Single Ease Question (SEQ). Table 9 shows the participants' assessments of LIMIS.

The range of the participants' answers was wide for most of the questions. P2, the most experienced radiologist with over 7 years of experience in annotating medical images, rated the system very favorably and liked the "novelty of [the] segmentation approach with text". Although P1 rated LIMIS with high values for effort and frustration, the participant stated that "once [...] [you] got into it, it was easy to use". Furthermore, the participants stated that the four predefined "suggestions are very valuable".

Table 9: Participants' answers to NASA TLX and SEQ.

	P1	P2	P3	P4
Mental Demand	14	5	11	8
Physical Demand	1	2	1	4
Temporal Demand	5	2	14	5
Performance	10	5	15	10
Effort	12	5	12	10
Frustration	14	1	18	10
SEQ	4	2	5	4

3.3.4 *Discussion on LIMIS*

We present LIMIS, the first language-only interactive model for medical imaging. Adapting a Grounded SAM-inspired architecture, LIMIS integrates problem-oriented multi-step language interactions with state-of-the-art medical foundation models, enabling accurate initial segmentations and user-driven mask adaptations. LIMIS was tested on multiple datasets, and its usability was evaluated by medical experts. LIMIS can be built on top of existing, promptable medical foundation models, which makes the approach flexible and, to some extent, model agnostic; however, it also ties the best-possible performance of LIMIS to the performance of the used foundation model.

3.4 CHAPTER CONCLUSION

In this chapter, we addressed the core challenge of creating large-scale anatomical datasets suitable for training deep learning models with full-body anatomical understanding. Manual annotations by medical professionals—while highly accurate, do not scale with the increasing data demands of modern AI systems, particularly medical foundation models. To overcome this, we proposed a novel automated pipeline for dataset generation, according to which we constructed the Dense Anatomical Prediction Atlas (DAP Atlas) dataset, a comprehensive dataset containing 142 anatomical labels across 533 full-body CT scans.

Our method builds upon public and private segmentation models, anatomical priors, and rule-based post-processing to derive a dense, holistic label map. These labels are refined through one iteration of self-training and are further corrected using anatomically informed post-processing rules. Importantly, DAP Atlas represents the first dataset to provide dense, voxel-wise annotations for the entire human body, from head to pelvis, without requiring direct manual annotation, while maintaining high anatomical fidelity.

We demonstrated the quality and utility of DAP Atlas through three validation strategies: (i) clinical expert evaluation, (ii) downstream segmentation performance on an independent benchmark (BTCV), and (iii) global anatomical plausibility checks. Each modality provided strong evidence for the reliability and practical value of the dataset. We developed a performant Atlas prediction model, which achieved a mean Dice of 85% on BTCV, matching state-of-the-art models trained in a fully supervised setting. Expert evaluation further emphasized the dataset's impressive quality, especially for educational and clinical visualization purposes, while also identifying systematic limitations in fine structures.

Finally, we analyzed the anatomical plausibility of the labels across demographic factors (e.g., age and sex) and demonstrated that the volume and intensity trends align with established medical knowledge. By leveraging the tumor metadata from AutoPET, we demonstrated the dataset's potential for pathology-related anatomical studies, making DAP Atlas not only a tool for segmentation but also for population-level anatomical and disease modeling.

The DAP Atlas project exemplifies a shift in dataset creation philosophy, moving from manual, small-scale, and local annotations towards automated, large-scale, and globally consistent anatomical mapping. This work provides both a blueprint and an openly available resource for the community, enabling future research in anatomical modeling, pathology integration, and clinically relevant AI applications.

With the increasing rise of large-scale datasets [130, 177, 178, 256, 335] we observe that label quality is an issue affecting all dataset creators as it becomes increasingly difficult to measure annotation quality, given the sheer amounts of automated [130] and semi-automatic [177, 333] masks in thousands of scans, which poses a central question for large-scale dataset creation: How good is good enough? In Section 3.2, we explore the trade-off between label quality and segmentation performance. While high-quality annotations are essential for small-scale datasets and in-domain tasks, they are costly and scale poorly. Through extensive experiments, we analyze how different degrees of label quality, ranging from weak pseudo-labels to near-expert segmentations, affect model performance across in-domain and pretraining tasks. Surprisingly, our findings suggest that for many scenarios, moderate label quality may suffice, challenging the notion that ever-improving annotations are always necessary. In summary, we find that for pretraining, it is not worth improving the label quality unless substantial gains can be achieved—minor improvements do not translate into better performance. This challenges the prevailing assumption that higher-quality annotations are always beneficial. Instead, our results suggest that the marginal utility of label quality diminishes with dataset size and that even noisy annotations can suffice for pretraining tasks. This raises several important questions for the field: Should efforts in dataset curation shift from precise annotation toward scalable label generation when training general-purpose models? To what extent can noisy annotations be tolerated before model performance degrades in real-world applications? And finally, do these findings hold for tasks beyond anatomical segmentation, also in pathology segmentation scenarios, where subtle structural changes are clinically significant and annotation precision may be paramount?

While our previous chapter demonstrated that label quality may not always be critical, especially in pretraining scenarios, real-world clinical use often demands reliable, case-specific segmentations. When predictions from automated models fall short, either due to noisy training labels, domain shift, or nuanced clinical requirements, post-hoc corrections become essential. To address this, we introduce LIMIS [105] in Section 3.3, a novel framework that enables intuitive, language-based interaction with segmentation models, allowing users to refine or generate masks in hands-free, high-stakes environments. LIMIS is evaluated on multiple datasets and tested with clinicians.

Contribution 1: We introduce the DAP Atlas dataset, a densely annotated whole-body CT dataset with 142 anatomical labels across 533 volumes, generated by aggregating public and private segmentation models, anatomical priors, and rule-based refinements in an automated fashion. We demonstrate the high anatomical fidelity and utility for downstream tasks through expert evaluation, external benchmark performance, and demographic plausibility analyses, establishing the DAP Atlas dataset generation approach as a scalable alternative to manual annotation.

Contribution 2: We systematically quantify the effect of label quality in medical segmentation by training models on datasets with varying pseudo-label accuracy and evaluating their performance across in-domain and pretraining settings. Our findings indicate that while label quality is crucial for in-domain segmentation, with diminishing returns in large-scale datasets, its influence on pretraining is almost negligible. Consequently, the information extracted during the pretraining phase appears to be more general concepts rather than specific details.

Contribution 3: We present LIMIS, the first language-based interactive medical image segmentation framework, enabling users to generate and, more importantly, refine unsatisfactory segmentation masks solely via natural language. The LIMIS approach is built on top of promptable medical foundation models and generates an initial segmentation from language by adapting the recent open-set object detectors to the medical field. We demonstrate the effectiveness of LIMIS across multiple datasets and in a user study with professionals.

4 LEVERAGING ANATOMICAL KNOWLEDGE FOR PATHOLOGY SEGMENTATION

Anatomical knowledge has the capability to enhance the segmentation of pathological structures by guiding models toward plausible regions, much like radiologists use anatomical context to resolve ambiguities and pinpoint abnormalities as diseases. We explore this property within this chapter by developing the Anatomy-Pathology Exchange (APEx) architecture in Section 4.1, where anatomical and pathological knowledge are jointly learned and evaluated in the CT and X-ray domains. While we employ the established joint-training procedure for the APEx architecture, we hypothesize that anatomical knowledge can also benefit pathological segmentations without being explicitly learned, but rather by leveraging anatomical knowledge that is present in existing anatomy segmentation models. In Section 4.2 we introduce the GRASP framework (Guided Representation Alignment for the Segmentation of Pathologies) by leveraging recent developments in anatomy segmentations, and building on top of our DAP Atlas dataset from Section 3.1.

The identification of pathological tissue in radiological scans is of key interest to the medical field, particularly in oncology, as it enables early diagnosis, monitoring of disease progression, and assessment of treatment response. Accurate delineation of such tissue is essential for clinical decision-making and can guide interventions such as surgery, radiotherapy, or personalized cancer-targeting therapies.

Within this chapter, we explore strategies to equip pathology segmentation models with anatomical understanding.

4.1 APEX: AN ANATOMY-PATHOLOGY KNOWLEDGE EXCHANGE ARCHITECTURE

The following section is based on our MICCAI 2024 work [128].

Throughout their extensive training, radiologists acquaint themselves with human biology and physiology, enabling them to discern typical patterns in the anatomy of both healthy individuals and those presenting health concerns. Years of clinical practice empower doctors to apply this underlying knowledge of the body to accurately associate subtle visual anatomy abnormalities with specific diseases. This holistic approach of doctors, considering both anatomy and pathology in the tissue, is contrasted by the vast amount of current automatic pathology segmentation models that specialize in narrow disease types and fall short of an overall understanding of body structures [216, 279]. These models are generally end-to-end semantic segmentation learners [240, 268], and resemble models designed for the natural image domain and as such could be applied interchangeably in both domains, from pathology- to street-scene- [369] and everyday object segmentation [75, 323]. Conversely, the medical imaging field has an obvious, yet often disregarded context: The human body.

While patients' anatomical features vary, the medical biases that associate anatomy with pathology for radiological assessment remain constant, such as simple observations, that a fracture has to be associated with a bone structure or that tumor locations often correspond to specific anatomical regions. When identifying a pathology, current segmentation models might or might not pick up anatomy-pathology correlations during training, which is the reverse direction to using anatomical priors for pathology identification. In the spirit of a doctor's workflow, we ask: *Can explicitly learned human anatomy improve a model's capability to predict pathological structures?*

In the following section, we explore various strategies to incorporate anatomical knowledge, represented by anatomical labels, to enhance pathology predictions. Inspired by the training of medical professionals, we propose a joint training procedure in which our network learns to predict both anatomy and pathology via our proposed architecture.

4.1.1 *Methodology*

We first present the learning setup for anatomy and pathology segmentation and walk through our ablations to incorporate anatomical knowledge into the model training. Finally, we derive our so-called **A**natomy-**P**athology **Ex**change (**APEx**) strategy to jointly learn both anatomy and pathology.

4.1.1.1 *Preliminaries*

Our formulation of the anatomy and pathology segmentation task depends on a training dataset:

$$\mathcal{D} = \{(x_i, a_i, p_i)\}_{i=0}^{N} \ , \tag{5}$$

with $x_i \in \mathbb{R}^{3\times H\times W}$ referring to one of the N images in the dataset, while $a_i \in [0, \dots, A]^{H\times W}$ is the associated anatomy with A classes and $p_i \in [0, \dots, P]^{H\times W}$ the pathology mask with P classes within the image. The task of a trained model is to predict, for new unseen test images x_t for each pixel in the image the correct anatomy categories a_t as well as the correct pathology classes p_t.

Table 10: Validation scores on the 5-fold CV PET/CT splits. A. Cond, A. Pred, and γ denote anatomy conditioning, auxiliary anatomy learning, and a weight factor, respectively.

Naive Anatomy Incooperation				
Method	A. Cond	A. Pred	γ	IoU
Baseline	–	–	–	54.34 ± 1.46
Pretrain	✓	–	–	56.64 ± 3.06
Multitask	–	✓	1	56.10 ± 3.36
Multitask	–	✓	10	57.12 ± 4.17
Multitask	–	✓	142	55.89 ± 3.03
Ana In	✓	–	–	57.23 ± 2.71
Ana In	✓	✓	1	56.52 ± 4.14

Architecture Ablations	
Method	IoU
Baseline	54.34 ± 1.46
+Shared BB	54.44 ± 4.14
+Shared PD	58.69 ± 3.63
└Query Sum	59.56 ± 3.64
└Query Sum 2-ways	59.35 ± 3.18
└Query Mean	59.78 ± 3.23
└Cross Attention (CA)	59.42 ± 2.42
└CA per feature level	58.48 ± 2.52

If the dataset provides instance-level annotations, we extend the approach to an instance-aware regime. Each anatomical mask a_i and pathological mask p_i then includes not only class- but instance-aware targets.

To investigate whether anatomical knowledge aids in identifying deviations from expected anatomy, we will examine two different tasks in two distinct domains: semantic segmentation of cancer in PET/CT images and instance-aware segmentation of thoracic abnormalities in chest X-rays.

To accommodate these varied requirements, we opt for a 2D model due to the constraints of the X-ray domain and model the 3D PET/CT images as sliced 2D images. To address the differing demands of semantic and instance-aware segmentation, we align with recent advancements in segmentation literature [39, 49, 175, 381] which intertwine both semantic- and instance segmentation through the design choice of predicting high-dimensional query vectors, which, combined with pixel-wise embeddings, encode instance-wise segments in an image. These queries are then employed to classify each segment, encapsulating information about both the segment's class and its shape. As a starting point for the experiments, we choose a Mask2Former [49] architecture. Our chosen setup is flexible in the choice of image modalities and in the choice of segmentation tasks.

4.1.1.2 *Incorporating Learned Anatomical Knowledge: A roadmap*

To investigate how to incorporate anatomical knowledge into the model training, we perform several ablations in a five-fold cross-validation setting in the domain of PET/CT. We built upon the established DAP Atlas Dataset introduced in Chapter 3 and AutoPET [81]. Details about the datasets are provided in Section 4.1.2. The baseline comparison model is a Mask2Former [49] model trained only on pathological labels. We report the 5-fold Validation IoU scores of naive anatomy incorporation techniques in Table 10 (left).

First, we investigate the effect of pretraining on anatomy. This leads to an improvement of about 2.3%.

Multitask Prediction: Next, we compare to jointly learned features using a multi-task setting approach. We paste the pathological labels on top of the anatomical labels, predicting an additional class. Despite being suboptimal, since PET/CT pixels could be interpreted as both anatomy and pathology, depending on the context, this leads to a similar improvement as pretraining. However, treating pathology as just another class underestimates its significance. To address this, we apply a weighted loss with weight γ, amplifying the pathology class's importance by 10-fold and 142-fold to equate it with the 142 anatomical labels. The 10-fold increase yields positive results, whereas the 142-fold adjustment demonstrates the challenge of selecting an appropriate weight factor.

Anatomy as an Auxiliary Input: Inspired by atlas-based segmentation methods, we input anatomical labels along with the PET/CT image, mimicking an optimal anatomy atlas. Using this procedure, we receive similar results to those of the previous approach.

4.1.1.3 *Architecture Ablations:*

The preceding analysis highlights that while anatomical knowledge can enhance pathology prediction, its effective utilization is a complex process. Thus, in our second experimental series, we postulate that due to the inherent overlap between anatomical and pathological labels, a two-head prediction approach is beneficial.

Initially, only the ResNet50 backbone is shared between the two prediction heads, resulting in no major improvement. A critical adjustment involves the sharing of a PixelDecoder across both anatomical and pathological prediction tasks. This integration significantly boosts the performance, evidenced by a notable increase of over 4% in IoU. This enhancement underscores the PixelDecoder's role in generating pixel embeddings rich in anatomical and pathological information, marking it as a crucial element in our design reflecting the dual role of each pixel in this task.

Query Mixing Strategies: Ultimately, as we employ distinct transformer decoders for anatomical and pathological predictions, we probe the efficacy of information exchange mechanisms via query exchange. This reflects the possibility of a direct exchange of queries representing anatomical and pathological segments. We explore various strategies, including nonparametric mixing and more flexible communication strategies such as cross-attention. While almost all strategies lead to a positive effect, none of them shines out as a clear winner.

We conclude this section with the insight that a two-head prediction approach, with one head for anatomy and one for pathologies, leveraging shared pixel embeddings, is a crucial design choice. On top, enabling communication between the different decoders leads to a further performance boost. The best-ablated model performs about 5.44% better than the naive baseline model.

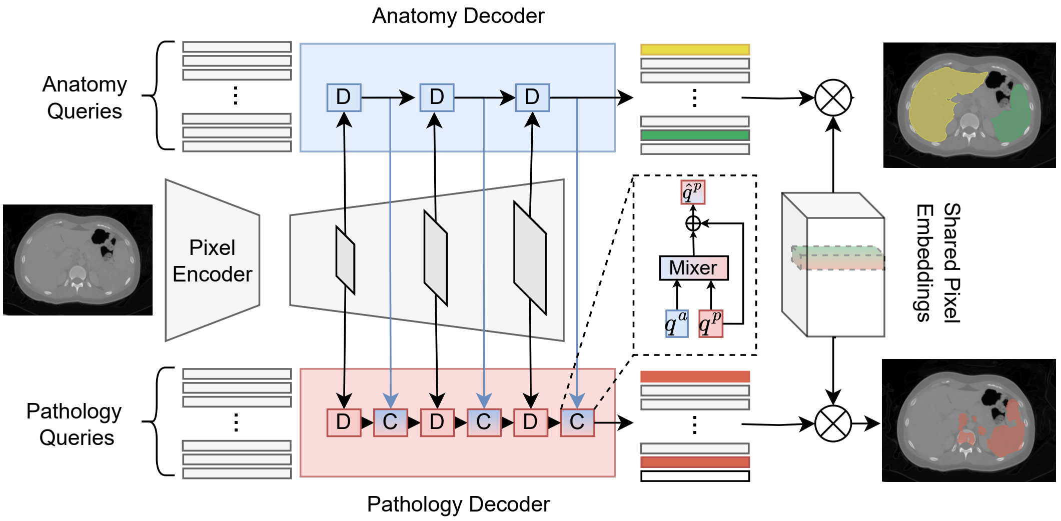


Figure 22: Overview of the proposed APEx Method, leveraging a shared pixel encoder, shared pixel embedding space, separate decoders, and a query-mixing module.

4.1.1.4 *Proposed Approach: APEx*

APEx is based on a query-based segmentation approach leveraging anatomical and pathological information. It incorporates anatomical context via the exchange of information between two decoders: One tasked to segment the anatomy and one tasked to segment the pathology. We show the overall method in Figure 22.

Shared Embedding Architecture:

Starting with a standard 2D image $x \in \mathbb{R}^{3\times H\times W}$ we encode the image using a feature extractor f^{extr} (parameterized by a ResNet50 [102]), which maps x to a set of feature maps at different scales $f^{extr}(x) = \{F_i\}_{i=0}^{n}$ with $F_i \in \mathbb{R}^{H_i\times W_i}$, such that $H_i > H_{i+1}$ and $W_i > W_{i+1}$ hold, i.e., feature maps successively get smaller in spatial extent. These feature maps are then decoded using an arbitrary pixel-decoder. We choose to use the deformable DETR [381] model as a pixel-decoder producing a set of enriched pixel embeddings $\{J_i\}_{i=0}^{n}$, with $J_i \in \mathbb{R}^{d\times H_i\times W_i}$.

Anatomy and Pathology Decoders:

Our architecture is motivated by computing separate query vectors for anatomy and pathology classes and letting the anatomy queries influence the pathology queries while limiting the reverse influence only to a shared embedding space.

Each enriched pixel encoding map J_i is accessed by two decoding functions $f_i^{ana}(\cdot)$ and $f_i^{path}(\cdot)$ from the function sets $\{f_i^{ana}(\cdot)\}_{i=n}^{1}$ and $\{f_i^{path}(\cdot)\}_{i=n}^{1}$ which either decode the anatomy or the pathology from it.

Randomly initialized, but learnable parameter-queries q_0^{ana} and q_0^{path} are transformed via

$$q_{i+1}^{ana} = f_i^{ana}(q_i^{ana}, J_i) \text{ and} \tag{6}$$

$$q_{i+1}^{path} = f_i^{path}(q_i^{path}, J_i) \ , \tag{7}$$

and optimized during training. The decoders $f_i(\cdot)$ follow a standard masked transformer setup, i.e. queries are transformed through a cross-attention layer that attends to the joint embeddings of the respective scale i, followed by a self-attention and feed-forward layer. For the pathology branch $\{f_i^{path}(\cdot)\}_{i=n}^{1}$ to explicitly adhere to the learned anatomical queries, an anatomy-to-pathology communication strategy is designed next.

Anatomy to Pathology Communication Strategy:

Medical personnel have access to a large amount of knowledge regarding the human body, which current pathology segmentation models do not have. Besides the implicit information exchange via the shared pixel embedding, we propose to integrate a communication step $f_i^{mix}(\cdot)$ after each pathology-decoder step $f_i^{path}(\cdot)$. There the queries q_i^{path} resulting from the scale i pathology-decoder are enriched with the anatomy queries q_i^{ana} from the anatomy-decoder as follows:

$$\hat{q}_i^{path} = f_i^{mix}(q_i^{ana}, q_i^{path}) \tag{8}$$

Here, $\hat{q}_i^{path}$ is the anatomy-enriched pathology query which, through a mixing strategy is capable of capturing anatomical information. We did not find a superior mixing strategy and thus would either recommend averaging the queries as a nonparametric approach or using a cross-attention mixing module.

In this asymmetric architectural setup, anatomical information influences the pathology-specific queries while the anatomy branch stays agnostic to any pathology and simply reflects the patient-specific anatomy details, serving as a useful foundational prior in pathology assessment. This design is ablated against an inferior design in which the anatomy branch is updated by the pathology as well (cf. Table 10: Query Sum 2-ways).

Joint Anatomy and Pathology Segmentation: Bringing the whole architecture and processing steps together into our Anatomy and Pathology Exchange (APEx) pipeline, we predict the anatomy and pathology segments through the following dot product:

$$out^{ana} = J_0 \cdot q_{n-1}^{ana} \text{ and} \tag{9}$$

$$out^{path} = J_0 \cdot \hat{q}_{n-1}^{path} \ , \tag{10}$$

Query vectors are passed through a simple classifier to associate anatomy or pathology classes to the predicted segments. The parameters of all components, namely f^{extr}, f^{ana}, f^{path}, and f^{mix} are optimized via weighted cross-entropy and binary mask losses enforced on each anatomy and pathology prediction out^{ana} and out^{path}.

Table 11: Comparison of APEx against multiple SOTA methods in the PET/CT domain (left). We highlight the **best** and the <u>second best</u> performance.

Method	PET/CT VAL			PET/CT TEST		
	IoU	Dice	BIoU	IoU	Dice	BIoU
DLV3+[43]	55.00 ± 3.5	70.91 ± 3.0	54.78 ± 3.6	53.60 ± 5.4	69.65 ± 4.8	53.07 ± 5.4
M2F[49]	54.34 ± 1.4	70.41 ± 1.22	54.16 ± 1.6	55.48 ± 1.1	71.36 ± 1.0	55.02 ± 1.1
UNET[268]	<u>57.62 ± 3.2</u>	<u>73.07 ± 2.6</u>	<u>57.38 ± 3.3</u>	<u>56.43 ± 1.5</u>	<u>72.14 ± 1.3</u>	<u>55.86 ± 1.4</u>
Ours (CA)	**59.43 ± 2.4**	**74.52 ± 1.9**	**59.21 ± 2.6**	**57.5 ± 0.9**	**73.01 ± 0.7**	**57.04 ± 0.9**

4.1.2 *Experiments and Results*

Datasets: To assess our method's broad applicability, we performed experiments across two vastly different medical imaging domains: FDG-PET/CT and Chest X-ray. For FDG-PET/CT, due to the absence of a comprehensive dataset with both anatomical and lesion annotations, we merged two distinct datasets: autoPET [81], which provides lesion annotations, and the DAP Atlas dataset [131], offering anatomical details. We exclude patients without pathologies, motivated by the high accuracy ($\geqslant 95\%$) of binary classifiers for cancer detection in PET images. Our study utilized 185 3D volumes for five-fold cross-validation and an additional test set of 125 cancer patients from the remaining dataset. To adapt images to the selected 2D setting, we slice them axially and stack CT and PET images channel-wise, leaving the third channel empty.

In the X-ray domain, we evaluate the properties of our method on the ChestXDet[186] dataset containing 13 pathology classes. To train anatomy segmentation, we predict anatomy pseudo-labels onto this dataset using a model trained on the PaxRay++ dataset [280]. We evaluate the different methods using five-fold cross-validation on the training set. During training, we omit images with no pathologies.

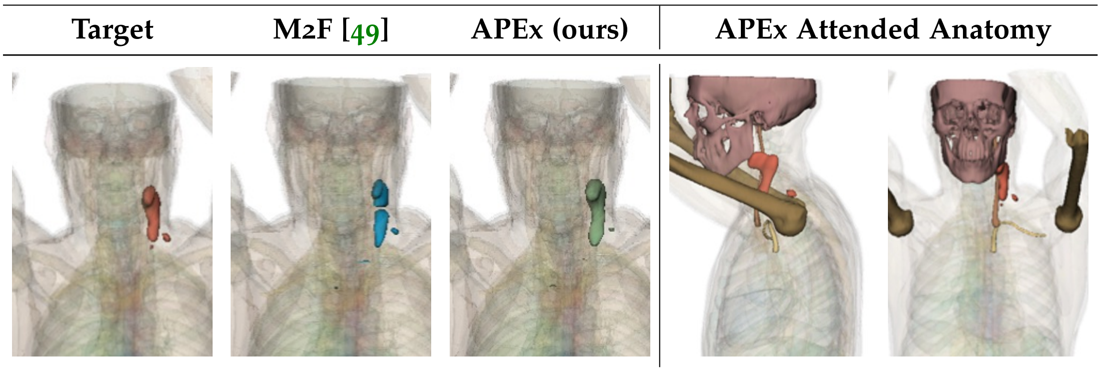


Figure 23: Stacked 2D tumor predictions next to top-5 attended anatomical structures.

Baselines and Methods: When evaluating models across different domains, we determine the best-performing candidate based on the performance on the individual validation sets. We use either the official test splits, if they exist, or a test set that we reserved beforehand. We benchmark APEx on PET/CT against established 2D segmentation baselines such as UNet [268], DeeplabV3+ [43], and Mask2Former [49]. In all experiments, we ensure models are trained using identical data and learning pipelines to isolate the effect of incorporating anatomical knowledge. For chest X-ray, we compare on instance segmentation against PointRend [155], MaskDino [175] and Mask2Former[49]. Regarding the specific APEx architecture, we choose the Cross-Attention Query Mixer, as it offered a competitive performance with the lowest standard deviation during our initial ablations (cf. Table 10).

4.1.2.1 *Semantic- and Instance Segmentation Results*

PET/CT Results: In Table 11 we report the Dice, IoU and Boundary IoU [47] (BIoU) performances of the previously mentioned baseline segmentation models against our method. All models have been initialized with LVM-MED weights [217] to provide a fair comparison. The results indicate that our method is capable of outperforming multiple strong competitors on our five-fold validation splits and the holdout testset. Figure 23 shows qualitative results as well as the most attended anatomical structures during the cross-attention query mixing step. We report a qualitative example in Figure 24. In the left part of Figure 24, we compare predictions of the baseline Mask2Former model against our developed APEx solution. While the baseline struggles to outline the tumor in the head-neck region correctly and misses a central part, APEx correctly predicts the entire metastasis. On the right side, we visualize which anatomical labels are attended by the pathology branch to inform the prediction of the displayed tumor. We visualize the anatomical labels alongside the ground-truth tumor mask in a sagittal and coronal view. We determine which anatomies are most relevant for the pathology branch by analyzing the attention matrix in the query mixer and identifying the anatomical structures that contributed the most to updating the pathology query. For the given example, these are mainly the *artery common carotid left*, the *artery subclavian left*, and three bone structures, namely the two *humeri* and the *skull*. We observe that a typical behavior is that the most relevant anatomical structures are either spatially close to the tumor or large bone structures, which are roughly in the same body region.

In addition to the previously reported results, we present qualitative examples in the PET/CT domain for intra-slice (Figure 40) and stacked predictions (Figure 24 & Figure 41) comparing Mask2Former with APEx. Within slices, APEx produces more precise structural delineations, while across slices, it misses fewer lesions and thus better captures the volumetric continuity of pathological structures, despite being trained exclusively on 2D slices.

Table 12: ChestXDet Results. We highlight the best performance in **bold** and the second best by underlining.

Pathology	MRCNN [101]	Casc. MRCNN [36]	PointRend [155]	MaskDino [175]	M2F [49]	Ours
Atel	3.92 ± 1.15	3.97 ± 0.90	4.22 ± 1.20	2.77 ± 0.77	3.83 ± 0.60	**4.38 ± 0.476**
Calc	5.37 ± 1.72	6.39 ± 1.89	**6.72 ± 1.75**	6.29 ± 0.86	5.62 ± 0.91	5.80 ± 1.39
Card	44.47 ± 2.51	45.82 ± 2.15	**50.03 ± 1.7**	37.18 ± 1.22	34.24 ± 5.69	48.00 ± 1.63
Consol	16.87 ± 0.412	17.65 ± 1.04	17.74 ± 0.51	18.79 ± 1.01	19.06 ± 1.14	**21.32 ± 0.59**
D.Nod	10.89 ± 2.70	13.12 ± 2.54	12.93 ± 2.96	18.33 ± 3.15	18.72 ± 2.88	**25.82 ± 2.27**
Eff	8.56 ± 1.10	8.28 ± 1.27	8.29 ± 0.35	10.78 ± 1.09	12.33 ± 0.63	**11.94 ± 1.47**
Emph	44.47 ± 2.31	42.89 ± 2.40	42.46 ± 3.43	45.9 ± 4.38	45.30 ± 2.04	**54.94 ± 2.61**
Fib	10.48 ± 2.11	10.71 ± 1.32	9.76 ± 1.69	10.31 ± 2.13	11.23 ± 1.85	**12.55 ± 1.39**
Fract	7.65 ± 1.08	6.36 ± 1.46	**10.61 ± 0.59**	8.527 ± 0.845	7.28 ± 0.86	9.76 ± 0.502
Mass	11.59 ± 2.89	11.93 ± 3.07	**15.66 ± 1.73**	12.39 ± 1.86	8.03 ± 1.40	11.94 ± 1.75
Nod	5.72 ± 0.53	5.41 ± 1.00	**6.97 ± 0.79**	5.646 ± 1.47	5.14 ± 0.59	5.48 ± 0.688
Pl.Thick	4.32 ± 1.32	4.43 ± 0.74	**5.12 ± 1.23**	4.28 ± 0.73	3.70 ± 0.50	4.61 ± 0.66
Pneumo	4.10 ± 0.804	3.21 ± 1.03	6.40 ± 0.92	5.64 ± 1.65	5.84 ± 1.24	**6.93 ± 1.06**
mAP	13.72 ± 0.41	13.86 ± 0.70	15.14 ± 0.44	14.38 ± 0.74	13.87 ± 0.53	**17.20 ± 0.33**

ChestXDet Results: In Table 12, we show the performance of different state-of-the-art instance segmentation methods trained using the same backbone. We see that our method improves over the Mask2Former baseline by ~3.75% mAP. Across 12 of 13 pathologies, our method achieves the best, or second-best performance, improving over recent transformer architectures as well as established CNN models. We report a qualitative example in Table 13. While all models struggle to approximate the ground-truth, APEx comes closest.

4.1.2.2 *Conclusion*

We proposed a novel way of leveraging anatomical information to improve pathology segmentation and showed the efficacy of the general concept of anatomy-guidance in two different domains covering diverse anatomical structures and pathologies. Besides improved performance, our method APEx encourages the exchange of anatomical information to ensure pathology segments are informed by the patient's anatomy, aligning more with the workflow of doctors that developed over decades.

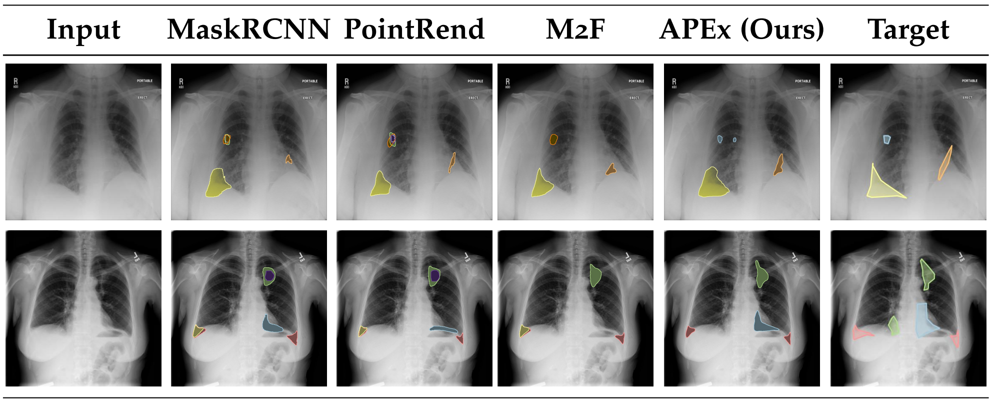


Table 13: Examples of qualitative segmentation results on ChestXDet. Each class is represented by a distinct color. Best viewed on screen with zoom and in color.

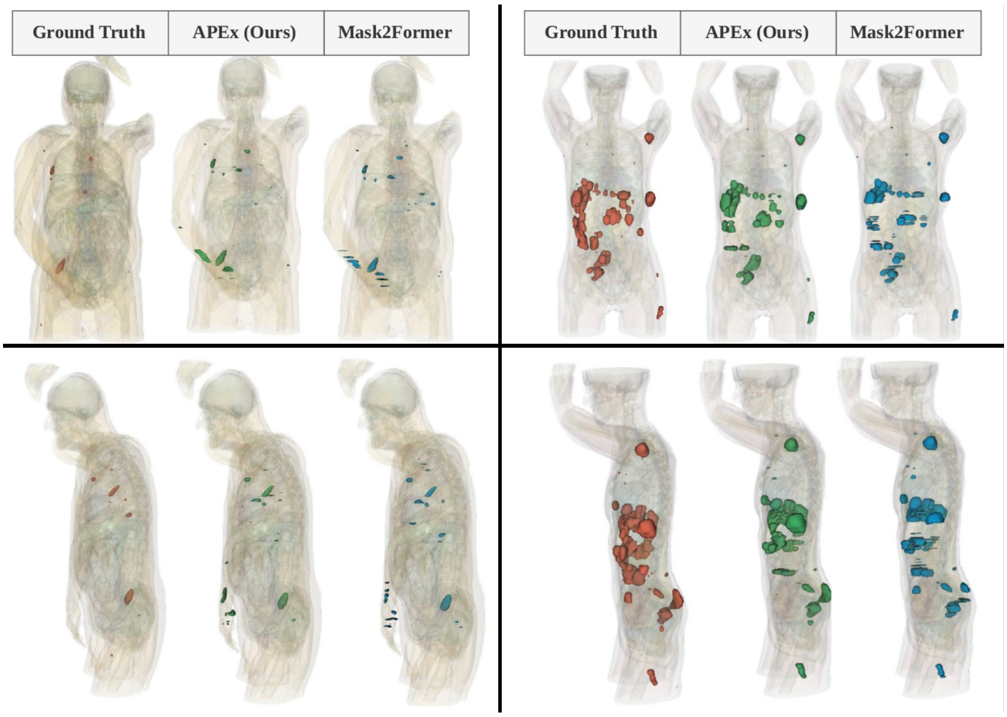


Figure 24: We show two patients in a coronal view on top and a sagittal view below. Volumes shown in red, green, and blue denote the lesion ground truth, APE predictions, and Mask2Former baseline predictions. Best viewed on a screen and in color.

4.2 GRASP: A PLUG-AND-PLAY ANATOMY-GUIDED SEGMENTATION FRAMEWORK

While the APEx architecture as outlined in Section 4.1 demonstrates that jointly learning anatomy and pathology can enhance segmentation performance, it requires modifying model architectures and additional anatomy supervision. Moreover, it is limited to a 2D formulation, adopted to enable experiments across both CT and X-ray domains within a unified framework. In the following section, we investigate whether anatomy-induced benefits can be achieved without explicitly retraining on anatomical labels by reusing existing anatomy segmentation models as frozen knowledge sources, an approach realized in our GRASP framework, which operates natively in 3D. This section is based on our MICCAI MLMI 2025 work [180].

Our approach is motivated by recent advances in holistic anatomical segmentation [103, 130, 333], where networks can capture large portions of human anatomy, yet anatomical and pathological segmentation remain largely separate domains. We thus ask the question: *Can we leverage recent advancements in anatomical segmentation models and their embedded anatomical knowledge to enhance pathology segmentation models, thereby aligning them more closely with human expert workflows?*

Recent work has shifted toward incorporating anatomical pseudo-labels directly into model training. APEx [128] as explored in Section 4.1, introduced dual-decoder joint segmentation for PET/CT and X-ray modalities, but operates in 2D, thus ignoring crucial volumetric interrelations. Multi-label approaches [229] have gained traction in AutoPET challenges, simultaneously predicting anatomical and tumor classes, but require careful organ selection and loss weighting to artificially focus on pathology classes [141], as standard losses lack inherent class preferences. Extensive multimodal pre-training with dual-decoder architectures [266] has shown improvements but demands complex dataset curation, loss balancing, and large computational requirements. These methods share common limitations: They require fundamental pipeline modifications or necessitate auxiliary training losses and extensive pre-training to learn anatomical representations from scratch. This raises a fundamental question: why reinvent anatomy segmentation as an auxiliary task when highly capable anatomical models already exist?

4.2.1 *Methodology*

We review a series of previously mentioned strategies to incorporate anatomical knowledge into the training of volumetric pathology segmentation models. We discuss their challenges before introducing the proposed GRASP framework. For these preliminary experiments, we use the 3D-UNet [53] architecture due to its popularity and experiment on the AutoPET [81] dataset, with PET/CT input $x := (x^{ct}, x^{pet})$.

4.2.1.1 *Exploratory Strategies for Anatomy-Pathology Alignment*

Transfer via Fine-Tuning: Fine-tuning naturally arises as a first approach when aiming to leverage anatomical knowledge for pathology segmentation. It offers a simple way to reuse spatial and semantic representations from well-pretrained anatomy models. Specifically, we fine-tune a pretrained anatomy model $f_{\theta_{\text{ana}}}$, originally trained on multi-organ labels $y^{\text{ana}} \in \{0, 1, \dots, \mathcal{C}_{\text{ana}}\}$ from the *DAP Atlas* [130], with $\mathcal{C}_{\text{ana}} \gg 2$. These rich anatomical features are expected to benefit the target binary task ($y^{\text{path}} \in \{0, 1\}$) despite the domain gap. We initialize $\theta \leftarrow \theta_{\text{ana}}$ and fine-tune on the pathology dataset $\mathcal{D}$:

$$\theta^*_{\text{path}} = \arg\min_{\theta} \sum_{(x_i, y_i^{\text{path}}) \in \mathcal{D}} \mathcal{L}(f_\theta(x_i), y_i^{\text{path}}). \tag{11}$$

The pretrained anatomy model is typically trained with CT-only input [130, 333]. Accordingly, we retain the pretrained weights for the CT channel and randomly initialize the PET channel. We consider two fine-tuning configurations: 1) a full 300-epoch schedule using the same learning rate as the baseline, and 2) a shorter 100-epoch schedule with a reduced learning rate. We report the results in Table 14 and observe that both approaches lead to a degraded performance compared to training from scratch. While initially surprising, this confirms similar results [141] and likely reflects the substantial class distribution shift. This underscores a key limitation: meaningful transfer requires careful and diverse dataset curation [266], which complicates the development of simple, generalizable solutions.

Multi-Class Supervision Strategy: While fine-tuning enables the reuse of anatomical representations, it maintains a strict separation between anatomy and pathology supervision. To more directly incorporate anatomical knowledge into the learning process, we adopt a unified multi-class formulation with a shared label space. Specifically, we define $y^{\text{multiclass}} \in \{0, 1, \dots, \mathcal{C}_{\text{ana}}, c_{\text{path}}\}$, where the tumor class c_{path} is appended as the last index. Anatomical pseudo-labels are derived from *TotalSegmentator* [333], with fine-grained substructures (e.g., individual lung segments) consolidated into one label per anatomical region to reduce GPU memory consumption. To better reflect the special role of the pathology class, we experiment with different strategies to enhance pathology importance via three loss weighting strategies: *Standard* (uniform), *Tumor-Focused* (tumor upweighted), and *Patch-Aware*. The *Patch-Aware* loss dynamically adjusts class weights per patch: absent classes are downweighted, anatomical classes receive moderate weight, and the tumor class is assigned the highest weight, promoting effective learning from both anatomy and pathology. Training follows the optimization objective displayed in Equation (12). Anatomical and Pathological labels are combined into a single multi-class target $y^{\text{multiclass}}$.

$$\theta^*_{\text{path}} = \arg\min_{\theta} \sum_{(x_i, y_i^{\text{multiclass}}) \in \mathcal{D}} \mathcal{L}_{\text{weighted}}(f_\theta(x_i), y_i^{\text{path}}). \tag{12}$$

We observe in Table 14 that under the standard loss, the results degrade compared to the baseline model. Only when we modify the loss to significantly enhance the importance of the tumor class under the *Patch-Aware* Loss setting does this approach slightly outperform the baseline; however, this marginal gain comes at the cost of substantial manual tuning and increased training complexity.

Multi-Task Learning Approach: To avoid coupling anatomical and pathological supervision in the multi-class formulation, we investigate a dual-branch architecture with a shared encoder $E_{\theta_{\text{enc}}}$ and two task-specific decoders $D_{\theta_{\text{ana}}}$ and $D_{\theta_{\text{path}}}$ for anatomy and pathology, respectively. This design enables task-specific predictions by decoding shared encoder features into separate anatomical and pathological outputs, supporting multi-label predictions and thus organ-pathology composition at inference time. The model is trained to simultaneously predict anatomical structures and tumor regions using separate output heads. The ground truth consists of two label maps: $y^{\text{path}} \in \{0, 1\}$ for binary tumor segmentation, and $y^{\text{ana}} \in \{0, 1, \dots, \mathcal{C}_{\text{ana}}\}$ for anatomical segmentation, where $\mathcal{C}_{\text{ana}}$ is the number of anatomical classes. The corresponding predictions are $\hat{y}_i^{\text{ana}} = D_{\theta_{\text{ana}}}(E_{\theta_{\text{enc}}}(x_i))$ and $\hat{y}_i^{\text{path}} = D_{\theta_{\text{path}}}(E_{\theta_{\text{enc}}}(x_i))$. Training is guided by a joint loss function that balances the two tasks:

$$\mathcal{L}_{\text{task}} = \alpha \cdot \mathcal{L}_{\text{path}} + (1 - \alpha) \cdot \mathcal{L}_{\text{ana}},$$

where $\mathcal{L}_{\text{path}}$ is a binary segmentation loss for pathology, $\mathcal{L}_{\text{ana}}$ is a multi-class segmentation loss over merged anatomical pseudo-labels, and $\alpha \in [0, 1]$ controls the relative importance of pathology supervision. To ensure a meaningful linear combination via α, we normalize both loss terms by their means to match their magnitudes. Results and an α ablation are shown in Table 14. We find that this approach does outperform the baseline and previous approaches by a small margin. Interestingly, we observe that when the pathology loss weight is set too high ($\alpha = 0.95$), the optimizer tends to disregard the anatomical component, leading to a decline in performance for the pathology as well.

Overall, we find, that some of the explored approaches deliver minor improvements over the baseline model but require substantial parameter tuning. A key insight we find, is that all of these methods enforce auxiliary losses upon anatomical labels, but do have pathology segmentation as the primary goal, requiring a parameter-based importance increase for the pathology task. We thus raise the question: Can we remove the auxiliary anatomical training from the training process by leveraging well-performing anatomy segmentation models and thereby keeping the focus of the target model on pathology segmentation? In the following, we propose the GRASP framework as a solution.

Table 14: Evaluation of pathology segmentation performances on AutoPET with a 3D-UNet backbone. LR indicates the learning rate; Loss FN, the loss function.

Models	Epochs	LR	Loss FN	Dice (↑)
3D-UNet (baseline)	300	1e−4	Standard DiceCE	49.3
Fine-tuning based	100	1e−5	Standard DiceCE	36.6
	300	1e−4	Standard DiceCE	22.3
Multi-class based	300	1e−4	Standard DiceCE	45.0
	300	1e−4	Tumor-Focused DiceCE	46.3
	300	1e−4	Patch-Aware DiceCE	49.6
Multi-task based	300	1e−4	normalized DiceCE ($\alpha = 0.7$)	51.0
	300	1e−4	normalized DiceCE ($\alpha = 0.8$)	51.8
	300	1e−4	normalized DiceCE ($\alpha = 0.9$)	51.7
	300	1e−4	normalized DiceCE ($\alpha = 0.95$)	47.6

4.2.2 *The GRASP Framework*

GRASP builds on the insight that high-quality anatomical models already exist. We introduce a plug-and-play framework that injects anatomical knowledge during pathology training by leveraging frozen anatomy encoders through feature alignment, without relying on auxiliary anatomical training. In GRASP, the CT modality effectively serves a dual purpose: it is the input to the anatomy and the pathology model. We illustrate the overall framework in Figure 25.

Dual Injection of Anatomical Priors: Anatomical knowledge is integrated through two complementary mechanisms: 1) anatomical pseudo-labels as an auxiliary input channel, and 2) bottleneck-level feature fusion. We follow an anatomical label-as-input strategy similar to the strategy developed for the APEx architecture in Section 4.1.1.2 by introducing a third input channel $x^{ana} \in \mathbb{R}^{H \times W \times D}$, which encodes voxel-wise anatomical pseudo-labels. This approach is inspired by atlas-based segmentation strategies [319], where anatomical priors guide label propagation. By providing pseudo-labels as an auxiliary input channel, it effectively supplies the model with a voxel-wise anatomical prior, enabling it to detect pathology as deviations from expected structure.

We use x^{ana} from a third model (e.g., *TotalSegmentator*), though outputs from the frozen anatomy model are also viable. For deeper integration, we also align and fuse anatomical features at the bottleneck level. Let the frozen pretrained anatomy model process the CT input to produce bottleneck features $Z_{ana} = E_{\theta_{ana}}(x^{ct})$, while the pathology encoder extracts joint features from CT, PET, and anatomical pseudo-labels (x^{ct}, x^{pet}, x^{ana}), yielding Z_{path}. As the spatial shape of Z_{ana} and Z_{path} may differ, we apply an *Align Block* to adjust anatomical features. It consists of a pointwise convolution followed by adaptive spatial pooling, transforming Z_{ana} to match the dimensionality of Z_{path}. We then feed both Z_{path} and the aligned Z_{ana} into our proposed fusion module to perform feature-level integration guided by pathological features.

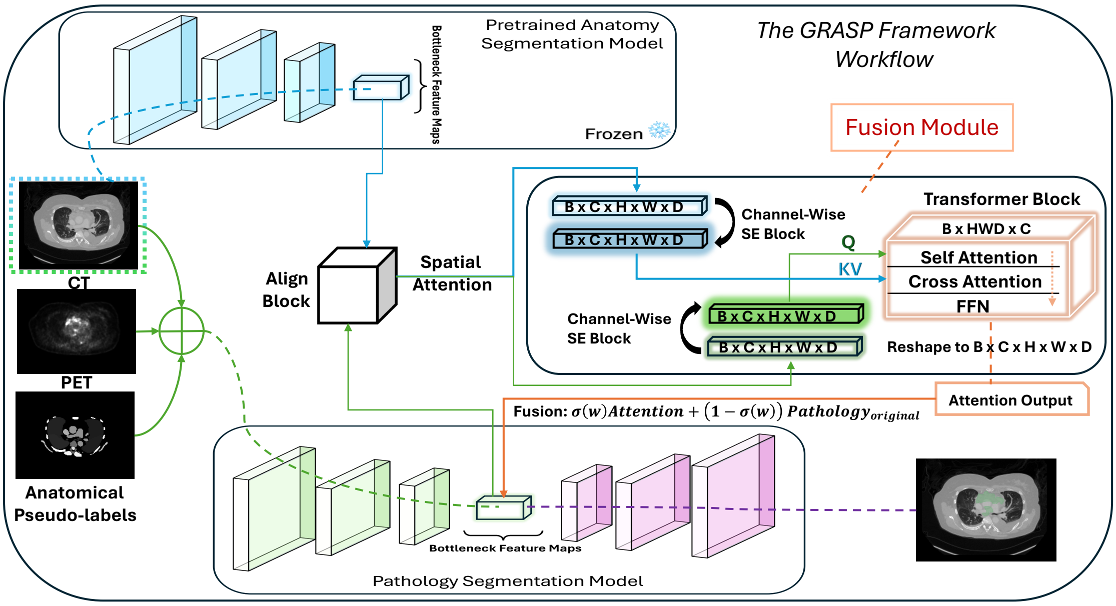


Figure 25: GRASP framework: Bottleneck features from a frozen pretrained anatomy model are aligned and fused with pathology features via a modular fusion block.

Anatomy-Guided Transformer Fusion: We design a lightweight fusion module leveraging transformer attention [312], as illustrated in Figure 25. We first apply a spatial attention (SA) block inspired by CBAM [339] to emphasize informative spatial regions. For each voxel location (h, w, d), attention weights are computed by aggregating feature responses across the channel dimension C via average and max pooling, followed by a convolution and sigmoid activation. Next, we apply a channel-wise Squeeze-and-Excitation (SE) [114] block to recalibrate the features, enhancing informative channels while suppressing less relevant ones. These two attention mechanisms are applied to both Z_{ana} and Z_{path}, yielding recalibrated features $\hat{Z}_{\text{ana}} \in \mathbb{R}^{B\times C\times H\times W\times D}$ and $\hat{Z}_{\text{path}} \in \mathbb{R}^{B\times C\times H\times W\times D}$, where B denotes the batch size:

$$\hat{Z}_{\text{ana}} = \text{SE}(\text{SA}(Z_{\text{ana}})), \hat{Z}_{\text{path}} = \text{SE}(\text{SA}(Z_{\text{path}})),$$

where $\text{SA}(\cdot)$ and $\text{SE}(\cdot)$ denote the spatial and channel attention modules, respectively. The recalibrated pathological features $\hat{Z}_{\text{path}}$ and anatomical features $\hat{Z}_{\text{ana}}$ are reshaped into sequences $Q = \text{reshape}(\hat{Z}_{\text{path}})$ and $K, V = \text{reshape}(\hat{Z}_{\text{ana}})$, all with shape $\mathbb{R}^{B\times(H\cdot W\cdot D)\times C}$. We first apply self-attention to Q to model intra-pathology dependencies, followed by cross-attention using Q as queries and K, V as keys and values. We adopt multi-head attention with $h = 8$ heads [312] to capture diverse anatomical-pathological relationships. The result is reshaped back and fused with the original Z_{path} via a learnable gated sum:

$$\tilde{Z} = \delta \cdot O(Q, K, V) + (1 - \delta) \cdot Z_{\text{path}}, \quad \delta = \texttt{Sigmoid}(w),$$

where $O(Q, K, V)$ denotes the output of the *Transformer Block*, with `Sigmoid`$(\cdot)$ controlling the contribution of the injected anatomical context, initialized to 0.5.
Fusion Setup and Training Strategy: We select up to the two deepest convolutional layers from the pathology backbone encoder as the bottleneck block for feature fusion, fusing each with the corresponding layer from the anatomy encoder in a pair-wise manner. To simplify the design, we propose two strategies: a *Mirror* setup, where a pathology model is trained from scratch and paired with a pretrained anatomy model of the same (mirrored) architecture trained on the *DAP Atlas* [130]; and a *Mixture* setup, where the same-architecture anatomy model is replaced with the off-the-shelf pretrained SegResNet [230] implemented by MONAI [38], improving generalizability. In practice, we employ a two-phase training strategy, where the fusion module is activated after 50 epochs, which has been shown to improve training stability.

4.2.3 *Experiments and Results*

Evaluation Protocol: We conduct experiments on two public PET/CT lesion segmentation datasets: AutoPET [81], a whole-body dataset with a high metastases count and HECKTOR [241], a dataset focusing on the head-neck region with fewer metastases. We benchmark GRASP using four complementary metrics: Dice (DSC) [61], CC-Dice (CC-DSC) [129], FP-Volume (FPV), and FN-Volume (FNV), computed per patient and averaged. For an easier comparison across the configurations, we compute the average rank across the four evaluation metrics for a final configuration rank. We experiment with three different segmentation models representing distinct architectural paradigms: 3D-UNet [53], the standard encoder-decoder baseline; SegResNet [230], a residual learning-based architecture; and MedNeXt-S [272], a modern and efficient ConvNeXt [194]-inspired design, based on our obtained pretrained models.
Implementation Details: Training uses AdamW [199] (lr=1e−4, cosine annealing [198]), batch size 4, 300 epochs, DiceCE loss, five-fold 70:30 splits. Patch sizes: (96,96,96) for AutoPET and (64,64,64) for HECKTOR. We use 2:1 positive-to-negative sampling, including healthy samples, unlike previous works [128, 141]. Our experiments are conducted on 4 NVIDIA H100 GPUs (80GB) with DDP. During inference, the feature fusion mechanism is omitted, relying solely on its regularizing effects on the decoder established during training. This approach facilitates easier model deployment and ensures that the feature fusion does not impose any additional computational burden during inference.
Quantitative Results: We report the quantitative results in Table 15, comparing baseline architectures against anatomical knowledge injection via a third input channel (ANA in.) and two GRASP variants (*Mirror* and *Mixture*). GRASP demonstrates strong performance, ranking first or second in nearly all configurations with only one exception. In a supplementary ablation on the 3D-UNet using the AutoPET dataset, we evaluated GRASP without providing anatomical pseudo-labels as additional in-

put channels, keeping only the feature fusion block. This variant improved Dice by +1.6% over the 2c baseline, but was still below the full 3-channel version, highlighting the complementary role of anatomical pseudo-label guidance and feature fusion. The framework shows strong adaptability across different anatomical segmentation architectures and consistently achieves the lowest FPV on AutoPET, effectively distinguishing metabolically active tissue from actual tumors. Performance on HECKTOR is also strong for the 3D U-Net and the SegResNet backbones. However, for MedNeXt-S, the simpler ANA in. approach slightly outperforms GRASP, suggesting that GRASP's more advanced fusion may be unnecessary for this particular backbone on this dataset. Overall, our results show that GRASP robustly improves segmentation across diverse medical imaging tasks by integrating anatomical knowledge.

Qualitative Results and Insights: We show qualitative results in Figure 26 (left) by exploring ground-truth lesions in purple against predictions of the established model configurations in red. We display two case studies representing difficult cases due to complex topology with many small lesions (Case A) and a complex surface structure (Case B). We further analyze the cosine similarity of pathology features before and after fusion. At epoch 50, when fusion begins, similarity drops sharply, indicating that anatomical feature inclusion rapidly shifts the pathology features. Both blocks show stabilization toward training's end, with pathology features maintaining 70-75% similarity before and after fusion, representing a 25-30% change due to anatomical feature inclusion.

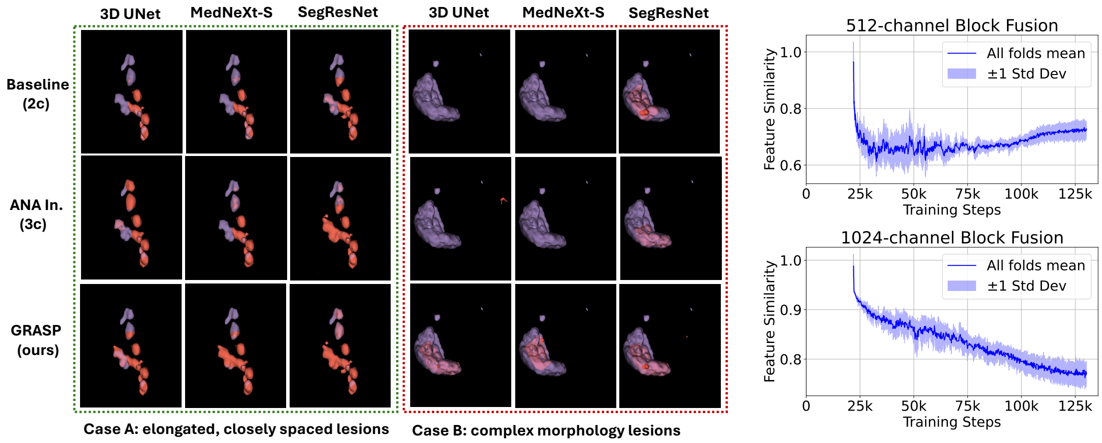


Figure 26: Left: Comparison of the three backbone model configurations on two cases, with ground truth in purple and predictions in red. Right: Feature similarity of pathological features before vs. after fusion in 3D-UNet's two fusion blocks.

Table 15: Benchmark segmentation results across two datasets. **Bold** indicates the best validation performance in each metric, while underline denotes the second-best.

Backbones	Configurations	DSC (↑)	CC-DSC (↑)	FPV (↓)	FNV (↓)	Rank (↓)
		AutoPET: High Metastases Count				
3D-UNet	Baseline (2c)	49.3 ± 2.9	31.4 ± 2.3	**2.49 ± 2.91**	29.96 ± 10.67	3
	ANA in. (3c)	52.6 ± 2.5	32.6 ± 2.6	3.16 ± 3.00	23.77 ± 5.52	3
	GRASP (Mixture)	53.6 ± 2.2	33.3 ± 0.9	2.80 ± 2.68	22.73 ± 6.09	2
	GRASP (Mirror)	**54.5 ± 3.0**	**34.7 ± 2.3**	2.77 ± 2.80	**19.75 ± 3.57**	1
MedNeXt-S	Baseline (2c)	50.3 ± 2.9	32.1 ± 2.1	3.95 ± 3.55	36.37 ± 13.42	4
	ANA in. (3c)	53.6 ± 3.3	**34.6 ± 3.0**	3.51 ± 3.23	27.99 ± 7.87	2
	GRASP (Mixture)	**53.8 ± 2.9**	34.1 ± 2.4	**2.91 ± 2.56**	28.54 ± 9.42	1
	GRASP (Mirror)	53.6 ± 3.2	33.7 ± 4.4	3.43± 3.04	**27.58± 12.40**	2
SegResNet	Baseline (2c)	54.1 ± 3.6	36.4 ± 3.5	4.08 ± 3.23	32.51 ± 13.74	3
	ANA in. (3c)	57.4 ± 3.6	37.2 ± 3.3	**3.39 ± 1.71**	21.89 ± 7.64	2
	GRASP (Mirror)	**58.9 ± 1.2**	**39.4** ± 2.3	5.87 ± 4.52	**17.08 ± 4.50**	1
		HECKTOR: Low Metastases Count				
3D-UNet	Baseline (2c)	39.9 ± 4.7	39.8 ± 4.7	**0.05 ± 0.03**	1.32 ± 0.26	4
	ANA in. (3c)	44.8 ± 3.9	44.6 ± 3.9	0.08 ± 0.04	1.06 ± 0.15	2
	GRASP (Mixture)	**47.1 ± 4.0**	**46.9 ± 4.0**	0.14 ± 0.11	**0.75 ± 0.15**	1
	GRASP (Mirror)	45.5 ± 5.5	45.3 ± 5.5	0.12 ± 0.06	1.16 ± 0.26	2
MedNeXt-S	Baseline (2c)	53.3 ± 3.2	53.2 ± 3.1	**0.27 ± 0.18**	0.55 ± 0.16	3
	ANA in. (3c)	**56.2 ± 3.3**	**56.0 ± 3.2**	0.39 ± 0.25	**0.44 ± 0.13**	1
	GRASP (Mixture)	54.3 ± 3.5	54.1 ± 3.4	0.34 ± 0.17	0.59 ± 0.14	3
	GRASP (Mirror)	55.2 ± 3.4	55.1 ± 3.4	0.38 ± 0.15	0.53 ± 0.20	2
SegResNet	Baseline (2c)	60.5 ± 2.9	60.4 ± 2.8	**0.24 ± 0.17**	0.48 ± 0.10	3
	ANA in. (3c)	61.9 ± 4.0	61.8 ± 4.0	0.68 ± 0.37	0.45 ± 0.11	2
	GRASP (Mirror)	**63.3 ± 3.0**	**63.2 ± 3.0**	0.43 ± 0.27	**0.35 ± 0.11**	1

4.3 CHAPTER CONCLUSION

Across both sections, we demonstrated that anatomical knowledge, if effectively integrated, can substantially enhance pathology segmentation. In APEx, we established that explicit joint learning of anatomy and pathology through a shared embedding space and targeted cross-decoder communication yields consistent gains across diverse modalities and tasks, aligning model behavior with clinical reasoning. However, this approach requires paired anatomy–pathology datasets, architectural modifications, and is restricted to 2D formulations.

GRASP, on the other hand, is a volumetric framework, decoupling anatomical learning from pathology training, instead exploiting frozen, high-quality anatomy segmentation models as external knowledge sources. This setup was motivated by the desire to avoid supervising anatomy as an auxiliary task when the anatomy output is irrelevant for the clinical application. Instead of learning anatomy from scratch, GRASP reuses frozen, high-quality anatomy segmentation models purely as knowledge sources. Anatomical priors are injected via pseudo-label input and bottleneck-level feature fusion, allowing the pathology model to remain fully focused on its target task while still benefiting from rich anatomical context. This design eliminates the need for paired labels, auxiliary losses, or architecture-heavy multi-task setups, and can be applied to volumetric architectures.

Together, APEx and GRASP provide complementary strategies: APEx excels when joint anatomy–pathology supervision is available, while GRASP offers a scalable, architecture-agnostic mechanism to harness anatomical priors from existing models. These findings reinforce the central role of anatomy-guided learning in advancing clinically relevant pathology segmentation.

Contribution 1: We show that incorporating anatomical knowledge consistently improves pathology segmentation across diverse tasks (semantic and instance), dimensionalities (2D and 3D), and imaging modalities (CT, PET, X-ray). This establishes anatomy-guidance as a generally applicable performance driver rather than a niche enhancement.

Contribution 2: We conduct a systematic exploration of anatomy–pathology integration strategies, ranging from pretraining and multi-class formulations to auxiliary-input and multi-task designs. This large-scale comparison reveals which mechanisms, such as shared representation spaces and selective cross-task communication, are effective, and which naive approaches fail to efficiently leverage anatomical priors.

Contribution 3: We distill these insights into two complementary frameworks: *APEx*, a specialized multitask architecture with asymmetric query-level information flow from anatomy to pathology, and *GRASP*, a plug-and-play approach that reuses frozen anatomy models for feature-level guidance without paired labels or inference overhead. Together, they provide practical solutions for both joint-training and anatomy-reuse scenarios.

5 EVALUATING LESION SEGMENTATION MODELS

In the preceding chapters, we first built a whole-body anatomy segmentation dataset and model in Chapter 3 and demonstrated how this anatomical knowledge, represented as anatomical labels, can enhance pathology segmentation in Chapter 4. In the following Chapter 5, we shift our focus to the evaluation of the trained lesion segmentation model. We critically assess existing semantic segmentation metrics which are used to measure the performance of lesion segmentation models, and develop a novel evaluation protocol: CC-Metrics, a framework that allows the evaluation of existing metrics on a per-component basis.

The selection of appropriate evaluation metrics is crucial in the development of neural networks, as these metrics directly influence model selection and optimization processes. A well-chosen metric ensures that the performance of a model represents the intended use case, thereby driving meaningful advancements in the field. Metrics that align closely with real-world applications help in fine-tuning models to meet specific needs, ensuring that improvements in performance translate into practical benefits. Thus, the careful consideration of evaluation metrics is fundamental to advancing the efficacy and reliability of neural networks.

5.1 EVERY COMPONENT COUNTS: REBALANCING SEMANTIC SEGMENTATION METRICS FOR MULTI-INSTANCE SCENARIOS

The following chapter is based on our work, published in AAAI 2025 [129].

Semantic segmentation [197] is a cornerstone of medical image analysis, as the automatic identification of critical areas, such as organs-at-risk [167] or metastases [81], can save valuable time in clinical care. With the ever-increasing performance of recent methods from the 3D-UNet [53], transformer-based models [98, 99] to the nnUNet [120], segmentation seems to be on the cusp of clinical use. When trying to translate these algorithms to actual clinical use, however, these models with high dice scores tend to produce irresponsible errors, such as the omission of novel, smaller lesions, which can significantly alter the treatment plan [30, 97]. The question thus becomes, how could such issues be identified in the development process before stress testing on patients?

In a typical setup, we aim to identify models that can predict both large and small structures while also maximizing the overlap between the predicted tumor regions and the actual tumors. This is non-trivial, as the selection of appropriate metrics for medical tasks depends on the specific scenario, the data at hand, the structure of the model's outputs, and the type of questions the researcher aims to answer. Recent publications highlight the pitfalls of using suboptimal metrics [262] and have developed extensive recommendation frameworks [209].

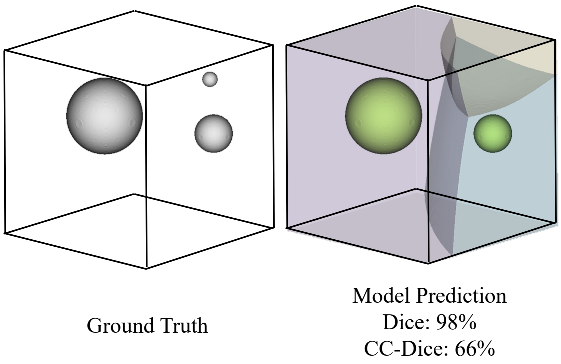


Figure 27: Reporting a Dice of 98% in the shown example highly overestimates the capability of the trained semantic segmentation model, possibly misleading radiologists. CC-Metrics partitions the image into distinct regions and evaluates standard segmentation metrics on a per-component basis, giving each tumor equal importance.

Despite the potential advantages of instance segmentation in distinguishing between overlapping objects, there appears to be less emphasis within the medical community on exploring and developing instance segmentation models for volumetric multi-instance segmentation scenarios, such as metastasis segmentation [81, 241].

We believe the main reason for this is: *Semantic Segmentation is sufficiently general*. Assume you have a perspective image of a street scene. Multiple objects and instances may overlap as the image is a projection of our three-dimensional world onto a two-dimensional plane. For instance, identifying individual people in a crowd does provide real value for downstream applications over a semantic-segmentation approach, where a crowd would only be represented as a blob. This scenario, by definition, is not possible in volumetric images, each connected component is perfectly separated, and there is no perspective overlap. A consequence of this observation is that standard semantic segmentation is sufficiently general to solve detection as well as instance segmentation, by computing connected components in a post-processing step. We will refer to this setup as "detection via segmentation".

Properly evaluating these semantic segmentation models in the context of a multi-instance scenario is challenging, as semantic segmentation metrics are inherently not designed to care about instances and compute the agreement of predictions and ground truth globally.

Within this work, we propose an embarrassingly simple yet intuitive strategy for evaluating the performance of semantic segmentation models in the 'detection via segmentation' setup by computing established semantic segmentation metrics on a per-instance basis. By doing so, we give equal weight to each component, reflecting their equal importance irrespective of their size. This approach aligns with the clinically motivated intent to treat all metastases the same.

To match predictions to ground truth values, we establish generalized Voronoi diagrams to partition each image into distinct regions, allowing predictions to be matched to the nearest ground truth connected component. By evaluating predictions locally, we eliminate the need for thresholds like those in Lesion Dice [225], allowing researchers to use existing, well-known metrics and avoiding the pitfalls of overlap-based matching or multiple true positives

5.1.1 *Overview of Existing Metrics*

Within this paragraph, we briefly revisit common metrics used to evaluate semantic segmentation models and point out related work that has raised criticism regarding the presented metrics within the medical field.

5.1.1.1 *Overlap-based Semantic Segmentation Measures:*

Overlap-based segmentation quality measures are one of the most frequently leveraged approaches to quantify the quality of segmentation masks. Typically, the predictions P are compared to the desired predictions S by their area of overlap. Two of the most common metrics used to quantify this overlap are the Jaccard Index [124], also called Intersection Over Union (IoU),

$$\mathrm{IoU} = \frac{|P \cap S|}{|P \cup S|} \tag{13}$$

and the Dice coefficient [61], which is defined in terms of set cardinalities as

$$\mathrm{Dice} = \frac{2 \times |P \cap S|}{|P| + |S|}. \tag{14}$$

A well-known limitation of these overlap-based metrics is their bias towards large objects and their inability to distinguish between different instances [262]. The latest research thus recommends reporting counting-based metrics (e.g., Precision) alongside overlap-based metrics [209]. To better counter size biases, a range of improvements, such as False Positive (FP) and False Negative (FN) penalties for the Dice score, have been proposed [37].

5.1.1.2 *Unified-measures for Segmentation and Recognition Quality:*

The Panoptic Quality (PQ) metric [153] was designed to unify detection and segmentation. PQ assigns predicted segments to ground truth segments by defining a match only if the predicted segment and ground truth segment overlap by at least IoU > 0.5, rendering a guaranteed unique matching. As outlined in Equation (15), the metric is calculated by computing the average IoU of all True-Positives (TP) and multiplying it by the F1-score [311].

$$\mathrm{PQ} = \frac{\sum_{(p,g)\in \mathrm{TP}} \mathrm{IoU}(p,g)}{|\mathrm{TP}|} \times \frac{|\mathrm{TP}|}{|\mathrm{TP}| + \frac{1}{2}\,|\mathrm{FP}| + \frac{1}{2}\,|\mathrm{FN}|} \tag{15}$$

The fixed-threshold-based matching approaches and direct metric combinations, like using the F1 score for counting and IoU for overlap, have limitations discussed in Section 5.1.4. Other works have also criticized the usage of the PQ metric in cell nuclei segmentation [71].

Recently, the Brats2023 challenge [225] started evaluating models based on a concept similar to Panoptic Quality by combining the overlap-based Dice score for all ground truth lesions, normalized by the number of all ground truth lesions and the number of FPs. They call this metric Lesion Dice (LD), and it is defined as follows:

$$\mathrm{LD} = \frac{\sum_{i\in(\mathrm{TP}\cup\mathrm{FN})} \mathrm{Dice}(i)}{|\mathrm{TP}| + |\mathrm{FP}| + |\mathrm{FN}|} \tag{16}$$

The assessment of TP, FP, and FN is, as in PQ, based on an overlap-based criterion. However, LD does not demand the strict IoU > 0.5 threshold which, as a consequence no longer guarantees a unique matching of prediction and ground truth.

5.1.1.3 *Boundary-based Semantic Segmentation measures:*

Boundary-based measures evaluate the quality of segmentations by focusing on the accuracy of the predicted boundaries relative to the ground truth. Two common metrics are Boundary IoU [47] and (Normalized) Surface Dice (NSD) [235, 281]. These metrics modify the traditional overlap-based metrics by considering only pixels within a specified distance from the boundary, emphasizing edge alignment. They are particularly useful in applications where boundary accuracy is crucial, such as pathology delimitation. Let ∂S and ∂P denote the surface voxels of the ground truth and prediction, then the NSD is defined as follows:

$$\mathrm{NSD} = \frac{|\partial S_\tau| + |\partial P_\tau|}{|\partial S| + |\partial P|} \tag{17}$$

where $\partial S_\tau = \{s \in \partial S \mid \exists p \in \partial P,\ d(s,p) \leqslant \tau\}$ is the set of surface pixels of the ground truth that are closer than a threshold τ to any of the surface pixels p of the predicted surface ∂P. ∂P_τ is defined analogously. For improved readability, we will use the term *Surface Dice* interchangeably with *normalized Surface Dice*, with both terms consistently referring to the normalized metric throughout this work.

5.1.1.4 *Distance-based Semantic Segmentation measures*

Distance-based measures quantify segmentation quality by evaluating the spatial distance between predicted and ground truth boundaries. The Hausdorff Distance (HD) [100] measures the maximum distance from a point on the boundary of one set to the closest point on the boundary of another set, highlighting worst-case boundary errors.

$$\mathrm{HD} = \max\left(\sup_{p\in P}\inf_{s\in S} d(p,s), \sup_{s\in S}\inf_{p\in P} d(s,p)\right) \tag{18}$$

As this measure is susceptible to outliers, researchers have started to report the 95th percentile, instead of the maximum distance which is known as the HD95 score. This worst-case behavior makes the metric less sensitive to small changes, especially for a large number of points [301].

Moreover, while the HD is commonly used in medical image segmentation, its application in pathology segmentation should be approached with caution. This metric emphasizes the maximum distance between boundary points, which can lead to inaccurate assessments, especially when used to detect multiple lesions as it defaults to the global worst case scenario.

Another popular metric is the Average Surface Distance [104] which computes the mean distance between corresponding points on the surfaces, providing an overall assessment of boundary alignment.

5.1.2 *Deriving CC-Metrics*

In this section, we introduce our proposed CC-Metrics evaluation protocol, focusing on the specific case of three dimensions ($d = 3$) due to its relevance in medical volumetric multi-instance semantic segmentation.

Consider an image $I \in \mathbb{R}^3$ and a binary target segmentation mask $S \in \{0,1\}^{h\times w\times d}$, where h, w, d denote height, width and depth respectively. Further, consider a neural network f computing binary predictions $P \in \{0,1\}^{h\times w\times d}$ given the image $f(I) = P$. To evaluate the quality Q of the prediction P, given the target segmentation mask S, the standard approach is, to use a metric m, taking both P and S and computing $Q = m(P,S)$. Typically, metric m generates a single quality measure, which globally measures some form of agreement between P and S. In the default setup, all predictions $p \in P$ and target segmentations $s \in S$ will be passed to the metric which results in the default global quality measure Q^{m}_{def}. In the following, we derive CC-Metrics by first introducing a generalized Voronoi diagram, which partitions the image space according to connected components of the ground truth S.

5.1.2.1 *Definition of the Generalized Voronoi Diagram*

Consider the standard definition of a Voronoi diagram in a discrete three-dimensional metric space $\mathbb{T}$ with distance function d, where $\mathbb{T}$ is a set of discrete points such as the indices of a tensor, or the integer lattice points in $\mathbb{R}^3$. Let $K \subset \mathbb{T}$ be a set of indices (3-tuples) which in our case can be thought of as a set of discrete locations in $\mathbb{T}$. To define the Voronoi diagram, a nonempty set of points $\{J_k\}_{k\in K}$ serves to define the sites of the diagram. Based on these definitions, a Voronoi Region R_k is the set of all points in $\mathbb{T}$ that are closer to $\{J_k\}_{k\in K}$, measured by distance function d than to any other site $\{J_l\}_{l\in K}$ with $k \neq l$. We define the distance function $d : \mathbb{T} \times K \mapsto \mathbb{R} : d(t, k) = \|t - k\|_2$ as a mapping from any location t in $\mathbb{T}$ and the location of any point $\{J_k\}_{k\in K}$ to their Euclidean distance. The standard Voronoi region R_k is thus defined as

Definition 1. *Voronoi Region*

$$R_k = \{t \in \mathbb{T} \mid d(t, J_k) < d(t, J_l) \quad \forall l \in K, l \neq k\}$$

This is the core concept according to which we plan to partition a given image tensor into regions. In our setup, the connected components of S will serve as the sites according to which the Voronoi regions are defined. As these components in most cases contain more than one single n-tuple (location) per Vornoi site, we expand the previous Definition 1 by allowing connected components to form the sites. We expand the definition of K by introducing $\mathbb{K}$ which bundles all points $\{J_k\}_{k\in K}$ into subsets based on their connectivity. Regarding the definition of connectivity, we rely on the concept of 26-connectivity. We define a function $\text{conn}_{26} : K \times K \mapsto \{\text{True}, \text{False}\}$, which takes the locations of two points J_{k_1} and J_{k_2} as inputs and determines whether they are 26 connected. The points J_{k_1} and J_{k_2} are connected if $k_1 = (a, b, c)$ and $k_2 = (d, e, f)$ meet the following condition:

Definition 2. *26-connectivity*

$$\begin{aligned}&|a - d| \leqslant 1 \wedge |b - e| \leqslant 1 \wedge |c - f| \leqslant 1,\\ &\textit{with } (a, b, c) \neq (d, e, f)\end{aligned}$$

Leveraging the established Definition 2, we can derive $\mathbb{K}$ which contains all sets of connected components according to the 26-connectivity criteria.

Definition 3. *Connected Component Sites*

$$\begin{aligned}\mathbb{K} = &\{C \subseteq K \mid \forall k_x, k_y \in C, \exists \text{sequence}(k_1, k_2, \ldots, k_l),\\ &\quad \textit{with } k_x = k_1, k_y = k_l \textit{ and } \text{conn}_{26}(k_i, k_{i+1}) \, \forall i < l\}\end{aligned}$$

This updated definition of Voronoi sites which we now call C reflects the generalized notion from single points to sets of connected points. We also modify the distance function to properly work on these sets of points, rather than two individual points. The updated function d' should map a location $t \in \mathbb{T}$ and a set of points $\{J_C\}_{C \in \mathbb{K}}$ to a distance. The intuition is, that each location should still be matched to its closest site. We thus define $d' : \mathbb{T} \times \mathbb{K} \mapsto \mathbb{R}$ as $d'(t, C) = \min_{k \in C} \|t - k\|_2$ as the minimal distance to any location k within the connecected component $C \in \mathbb{K}$. With the updated distance function, the generalized Voronoi region R' can be easily defined as

Definition 4. *Generalized Voronoi Region*

$$\begin{aligned} R'_C = \{ & t \in \mathbb{T} \mid d'(t, J_{C_k}) < d'(t, J_{C_l}) \\ & \forall C_k, C_l \in \mathbb{K},\ C_k \neq C_l \} \end{aligned}$$

This definition leads to a deterministic and unique separation which we proof in Appendix E.1.3

5.1.2.2 *Calculating CC-Metrics*

After establishing the definition of the Generalized Voronoi regions, the calculations of CC-Metrics is straight forward.

Let the current image I be defined over a metric space $\mathbb{T}$ with indices (a, b, c) for the 3D volume and $f(I) = P$ be the predictions of a neural network. The target segmentation mask $S \in \{0, 1\}^{h \times w \times d}$ can now be used to define the set of indices K as

$$K = \{(a, b, c) \in \mathbb{T} \mid S(i, j, k) = 1\}$$

In this definition, K includes all the indices where the target segmentation mask has the value 1. Given K, $\mathbb{K}$ can be computed using the conn26 function and R'_C can be computed using Algorithm 1 or Algorithm 2.

We now define the local predictions P_C for the region R'_C as $P_C = P \cap R'_C$, where P_C represents the set of predicted points in the Voronoi region R'_C. Similarly, we only consider the target segmentation within the same region $S_C = S \cap R'_C$. We now compute the metric of interest locally and separately for all regions

$$Q^m_C = m(P_C, S_C) \quad \forall C \in \mathbb{K}$$

ensuring that the evaluation is constrained to the specific region of interest.

We aggregate the different local quality measures using a standard average

$$Q^m_{glob} = \frac{1}{|\mathbb{K}|} \sum_{C \in \mathbb{K}} Q^m_C$$

5.1.3 *Computation of Generalized Voronoi Diagrams*

In this section, we discuss the algorithm to compute the Generalized Voronoi Diagrams R'_C. Algorithm 1 first computes connected components so that each connected component in the original mask S is distinguishable. We utilize the implementation by Silversmith [292] for this purpose. Subsequently, we compute the Euclidean distance transform for each component separately. We use the scipy implementation [313] for this computation. However, since the scipy implementation computes the distance transform of the pixels inside each component, and we are interested in the distances of the pixels of the background class, we invert the pixel values, setting everything except the current component to the foreground and the pixels within the current component to the background. As this step is specific to the library used, we omit it from the pseudo- algorithm 1.

We collect the individual distance transforms and stack them in order. To compute the generalized Voronoi Regions R' for the entire image, we take the argument of the minimal distance across all distance transforms. This operation assigns each voxel to its closest component as measured in Euclidean distance, which matches the definition of the generalized Voronoi Diagram.

The proposed Algorithm 1, while intuitive, is of time complexity $\mathcal{O}(|\mathbb{K}| * |S|)$ with $|\mathbb{K}|$ being the number of connected components and $|S|$ being the number of voxels in the segmentation mask S. To speed up the computation, we further develop Algorithm 2, which leverages a feature transform algorithm, thereby assigning the nearest component to every voxel in one pass.

To compute generalized Voronoi regions R'_C more efficiently, we exploit the feature transform available in the `scipy.ndimage.distance_transform_edt` function [313].

This approach is mathematically equivalent to the naive method (Algorithm 1), where a distance transform is computed separately for each component and the minimal distance is selected voxel-wise. However, the feature transform method (Algorithm 2) requires only a single distance transform computation. As a result, its runtime is $\mathcal{O}(|S|)$, still scaling linearly with the total number of voxels, but is now independent of the number of connected components, which leads to significant speedups in practice, particularly for high-metastases count datasets.

5.1.4 *Analysis of Segmentation Metrics*

Within the following section, we analyze failure cases of common semantic segmentation metrics used in a "detection via segmentation" scenario, based on a toy example. For all of the following analyses, we start, unless noted otherwise, with a ground truth consisting of three different sphere components. A visualization of this ground truth is given in Figure 27. Initially, we assume a perfect prediction which we then degrade step by step as described in the following sections.

Algorithm 1 Compute Generalized Voronoi Diagrams

```
Require: Segmentation mask $S \in \{0,1\}^{h \times w \times d}$
Ensure: Generalized Voronoi Regions $R'_C \quad \forall C \in \mathbb{K}$
  Step 1: Compute Connected Components
  $\mathbb{K} \leftarrow \text{LabelConnectedComponents}(S)$
  Step 2: Compute Euclidean Distance Transforms
  $\text{cc\_dt} \leftarrow [\,]$
  for each connected component C in $\mathbb{K}$ do
      $D_C \leftarrow \text{euclidean\_dist\_transform(background of } C)$
      $\text{cc\_dt.append}(D_C)$
  end for
  Step 3: Compute Voronoi Regions
  $R'_C \leftarrow \text{argmin(stack(cc\_dt))}$
```

Algorithm 2 Fast Computation of Generalized Voronoi Diagrams

```
Require: Segmentation mask $S \in \{0,1\}^{h \times w \times d}$, connectivity $c \in \{6, 18, 26\}$
Ensure: Voronoi regions $R'_C \forall C \in \mathbb{K}$
  Step 1: Label Connected Components
  $\mathbb{K} \leftarrow \text{LabelConnectedComponents}(S, \text{connectivity} = c)$
  Step 2: Prepare Distance Transform Input
  $I \leftarrow \mathbf{1}$                          ▷ Start with all voxels marked as background
  $I[\mathbb{K} > 0] \leftarrow 0$              ▷ Set voxels in connected components to zero (seeds)
  Step 2: Compute Nearest Seed Indices
  $\text{indices} \leftarrow \text{Feature Transform}(I)$
  Step 3: Assign Voronoi Regions
  $R'_C \leftarrow \mathbb{K}[\text{indices}]$
  return $R'_C$
```

5.1.4.1 *Overlap Metrics: Dice vs. CC-Dice*

First, we compare the Dice metric as an example of an overlap-based metric. As an operation to degrade predictions, we continuously apply binary erosion to the prediction masks. We show the results of two experiments in Figure 28 in the top and middle plots. In the upper plot, we uniformly apply erosion to all components. It's important to note that binary erosion removes a larger percentage of volume from smaller spheres than from larger ones. As a result, we observe that the individual Dice scores of three components (three black dashed lines) decrease with different speeds due to their different sizes. The global standard Dice coefficient (blue line) decreases more slowly and almost follows the Dice of the largest connected component. CC-Dice (orange line) decreases faster as it reflects the average per component Dice. In the second example (middle plot), we only apply erosion to the smallest component. The standard Dice is barely affected while CC-Dice quickly converges to 66% reflecting perfect coverage for 2 components and 0% Dice for the smallest component.

5.1.4.2 *Unified Metrics: Panoptic Quality vs. CC-Dice*

Panoptic Quality (PQ) [153] is a metric that measures both recognition and segmentation quality, making it suitable for comparison with CC-Dice.

We compare CC-Dice and PQ in Figure 28 (top and middle plot) under the previously described scenarios. We observe two characteristic weaknesses of PQ. The first weakness is evident at erosion steps 3, 7, and 14 in the upper plot, where PQ drops rapidly. At other steps, the decline is smoother. This occurs because, at these erosion steps, the three different segmentation components fall below the IoU > 0.5 thresholds relative to their ground truth, abruptly changing each component from TP to FP and causing sharp drops in the metric. In contrast, CC-Dice, which does not rely on fixed thresholds, behaves much more smoothly.

Another consequence of fixed thresholds is best observed in the middle plot of Figure 28. Since PQ only measures segmentation quality for TPs, the score remains flat once the smallest component is no longer accepted as a TP. PQ does not reflect changes in segmentation quality after falling below this threshold. On the other hand, CC-Dice remains informative even for low overlaps.

A second major negative property of PQ is the direct combination of counting-based and overlap-based scores. As mentioned earlier, once the IoU of a component drops below 0.5, it is no longer considered a TP but becomes an FP, which negatively impacts the metric. However, after the component is completely eroded, the PQ score increases again because the presumed FP component is no longer present. This behavior is observed at erosion step 12 in the upper plot and erosion step 20 in the middle plot. We find this behavior suboptimal, as it introduces inconsistencies and irregularities in the evaluation of segmentation quality. The abrupt changes in PQ due to the fixed IoU threshold can complicate network evaluation, as minor variations in detected components may drastically alter the score. Additionally, the increase in PQ after a component is fully eroded is suboptimal, as it rewards the absence of predictions rather than partially correct ones. This is problematic in cases where missed components are more harmful than False Positives. While not designed to compete with PQ, CC-Dice provides a more consistent and representative assessment of segmentation quality by avoiding these issues.

5.1.4.3 *Lesion Dice vs. CC-Dice*

Besides PQ we compare Lesion Dice (LD) as a second unified metric against CC-Dice.

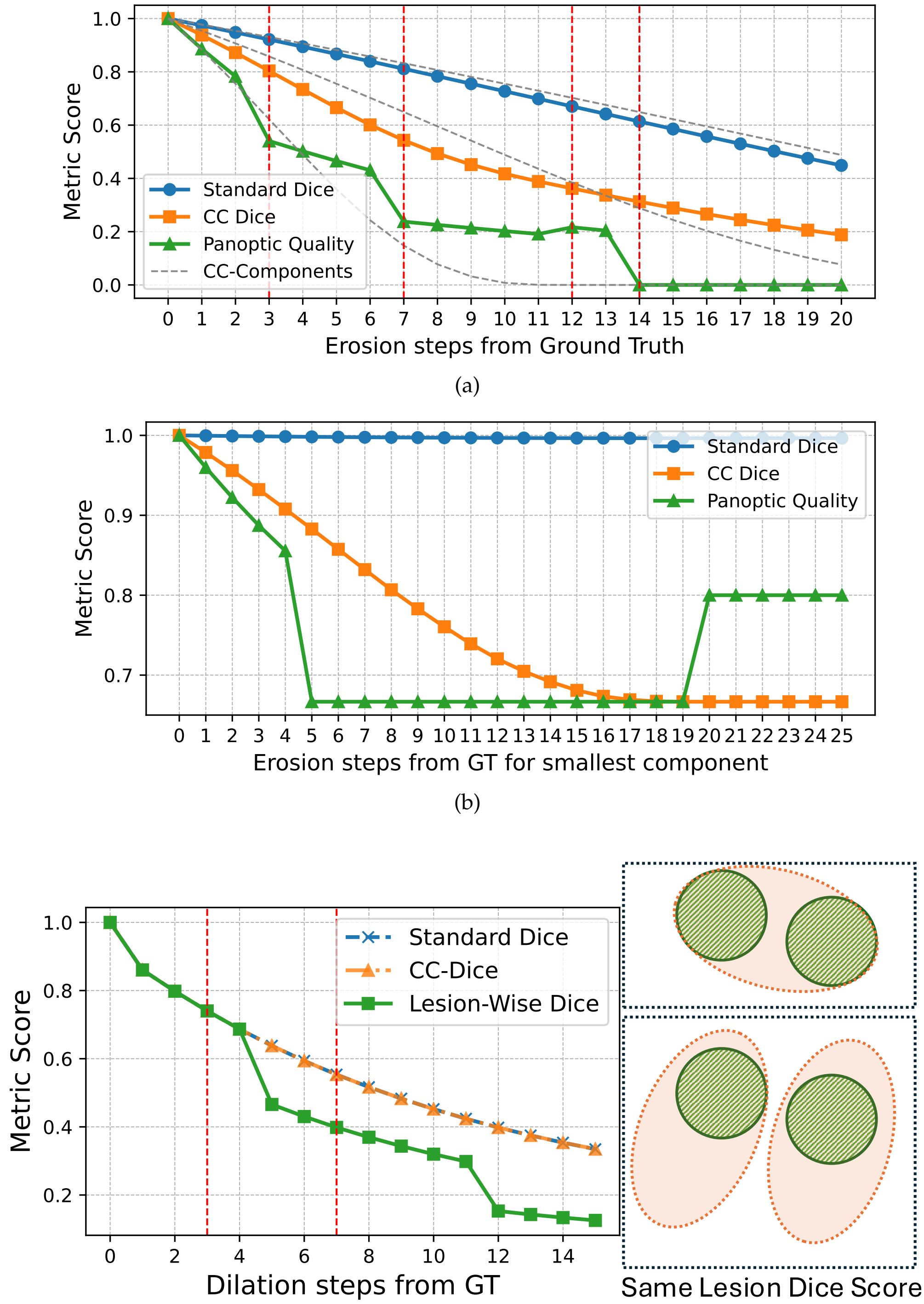
LD requires the careful selection of many domain-specific hyperparameters, such as dilating the ground truth n-times to merge adjacent ground truth instances or ignoring predictions with components smaller than k milliliters. Many of these parameters have to be chosen by experts [225]. While this adds flexibility, it places the burden of threshold selection on researchers and complicates cross-domain comparisons. In contrast, CC-Dice is a hyperparameter-free evaluation protocol.

Similar to PQ, LD directly incorporates the number of false positives into the score if they form components larger than k milliliters. This creates a challenge in selecting an appropriate k value. If k is set high, the metric may ignore numerous false positives below k without affecting the score. Conversely, if k is low, even a few false positive pixels can significantly lower the score, in this setting, we penalize the model for uncertain predictions in a setting where the cost of a false positive is much lower than that of a false negative.

Unlike PQ, Lesion-Dice does not demand an IoU > 0.5; even a single overlapping pixel can determine a match between prediction and ground truth. This no longer guarantees a unique matching, hence multiple masks with minimal overlap can be counted as TPs.

The lower plot of Figure 28 illustrates two pitfalls of LD. In the left graph, we start with three spheres of equal size and dilate the predicted masks. The Dice score (blue line) due to each component being of equal size behaves in line with CC-Dice (orange line). LD (green line) initially follows this pattern, however, once two previously not connected masks intersect, LS assigns the now connected mask as a TP to both lesions thereby decreasing the scores massively as seen at dilation steps 3 and 7.

The double assignment in LD also leads to unexpected results, as demonstrated in the pitfall example on the right side of the lower plot. LD assigns the mask as a TP to both components individually. This leads to the scenario that two separate masks of the same size are not preferred over a single mask covering both components, despite the two masks covering many more false-positive locations.



(a)

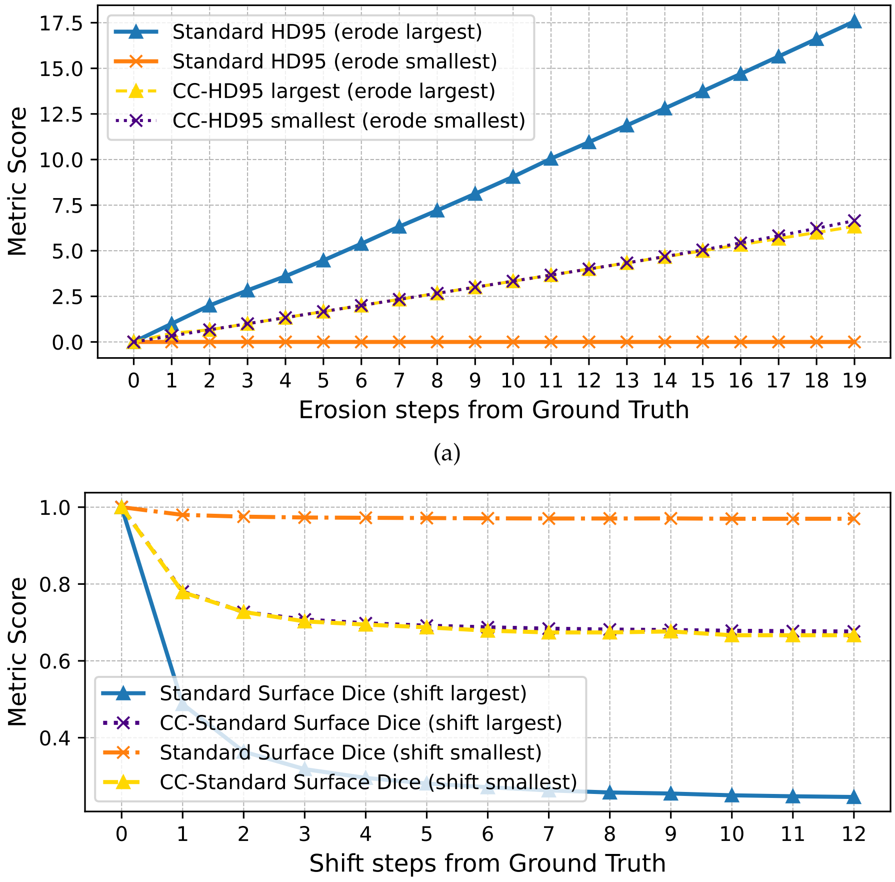
(b)

(d)

(e)

(e)

Figure 28: Comparison of Dice, CC-Dice and Panoptic Quality: In the upper plot we start from a perfect prediction and degrade prediction quality by applying erosion to all components uniformly. In the middle plot we only degrade the prediction of the smallest mask. In the lower plot, we compare CC-Dice with Lesion Dice (LD) by using dilation to simulate oversegmentation (left) and highlight a pitfall of LD (right).



(a)

(b)

Figure 29: Comparison of the standard Hausdorff95 metric with the CC-Hausdorff95 metric (upper plot), as well as standard Surface Dice with CC-Surface Dice (lower plot). In both scenarios, we start from a perfect prediction and assess the metric scores while degrading the prediction quality of a large versus a small component.

5.1.4.4 *Distance Metrics: Hausdorff Distance vs. CC-Hausdorff Distance*

Distance-based metrics measure the distance between the boundaries of two masks. While this approach can be executed globally, large components, which naturally contain the vast majority of surface points, limit the importance of smaller components, either because their distances are treated as outliers and ignored, as in Hausdorff95 distance, or count little towards the average in Normalized Surface Distance. We explore this behavior in Figure 29. Again, we start with the spheres displayed in Figure 27 and assume a perfect prediction. We then gauge how the HD95 metric changes during the erosion of the smallest segmentation mask compared to the erosion of the largest segmentation mask. It can be seen that, using the standard evaluation protocol, the small component is ignored because its boundary distances exceed the 95th percentile, whereas eroding the largest component results in a significant increase in the metric. Evaluating the HD95 distance on a per-component basis normalizes the different number of boundary pixels and only compares one ground truth mask to its respective prediction. As desired, the CC-HD95 distances behave almost identically whether we erode the largest segmentation mask or the smallest segmentation mask.

5.1.4.5 *Boundary-based Metric: Surface Dice vs CC-Surface Dice*

Boundary-based metrics suffer from the same problem as overlap-based metrics when being evaluated globally. Large components having long boundaries limit the influence of smaller components. This behavior can be observed in the bottom plot of Figure 29. In this example, we use a shift operation to degrade prediction quality, as erosion would leave the metric unchanged until the threshold τ in Equation (17) is reached, at which point it would drop to zero. When degrading the largest component (blue line), the Surface Dice score decreases significantly, while degrading the smallest component (orange line) barely affects the overall score. Evaluating on a per-component basis results in consistent scores for the same operations, regardless of component size.

5.1.5 *Evaluation*

To evaluate the proposed CC-Metrics protocol on PET/CT datasets, we conduct experiments on two publicly available datasets: AutoPET [81] and HECKTOR [241], ensuring a comprehensive analysis across different cancer types. These datasets were selected as AutoPET involves patients with an average of 10 metastases, allowing us to assess performance in complex scenarios, while HECKTOR includes patients with an average of only 2 metastases per image, providing a setting where we expect standard metrics and CC-Metrics to be aligned.

We first simulate a range of different model failures using synthetic predictions. Given that HECKTOR includes only a few metastases per patient, AutoPET, with its higher average number of metastases, provides a more suitable dataset for our simulation. This choice allows us to effectively gauge how CC-Metrics behave in comparison to the standard evaluation protocol on a dataset level. Next, we train segmentation models on both the AutoPET and HECKTOR datasets and evaluate them using both approaches, ensuring that CC-Metrics and standard metrics are assessed across scenarios with varying metastasis counts.

5.1.5.1 *Evaluation on Synthetic Prediction*

We initially assume a perfect prediction for each image and then apply a degrading function to progressively worsen and edit this prediction over n steps. At each degrading step, we aggregate predictions, considering patients with at least n metastases when altering n components.

The simulation results for the AutoPET dataset are shown in Figures 30 to 33. The first row of each scenario simulates decreasing prediction quality for the n smallest components, and the second row for the n largest. We report the median along with the 25th and 75th percentiles of standard metrics in orange and CC-Metrics in blue.

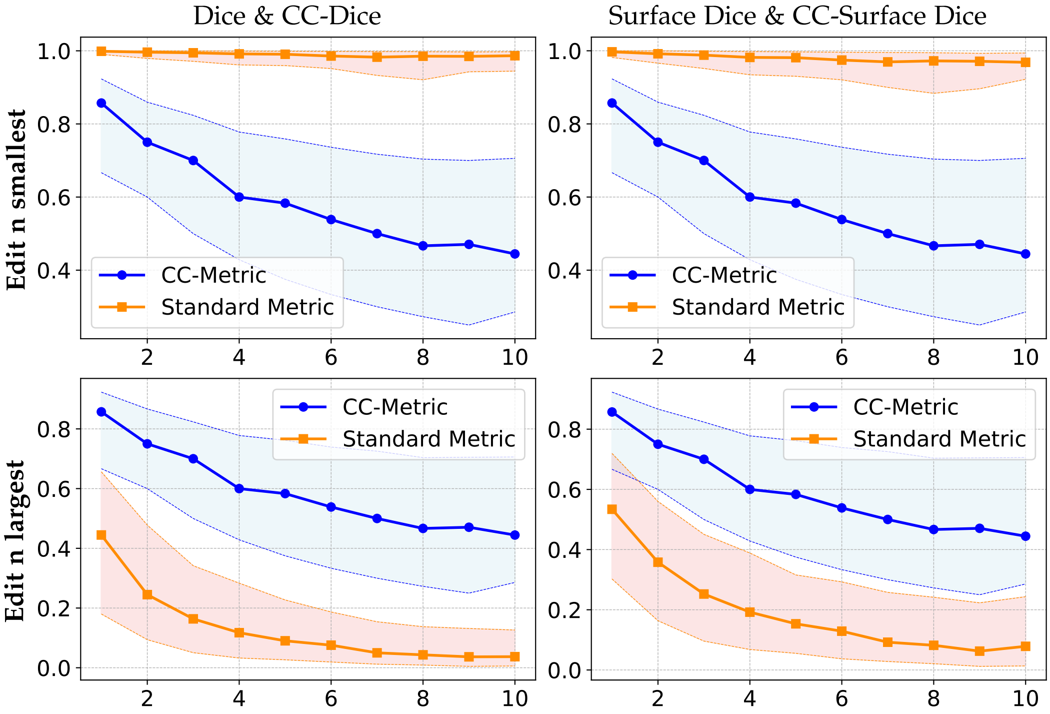


Figure 30: Comparison of standard and CC semantic segmentation metrics on the AutoPET dataset: We drop n connected components, simulating that lesions were forgotten by a model. Standard Metrics are shown in orange, CC-Metrics are shown in blue.

False Negatives (Figure 30): We simulate the occurrence of false negatives by dropping components. We observe that both standard Dice and Surface Dice are heavily

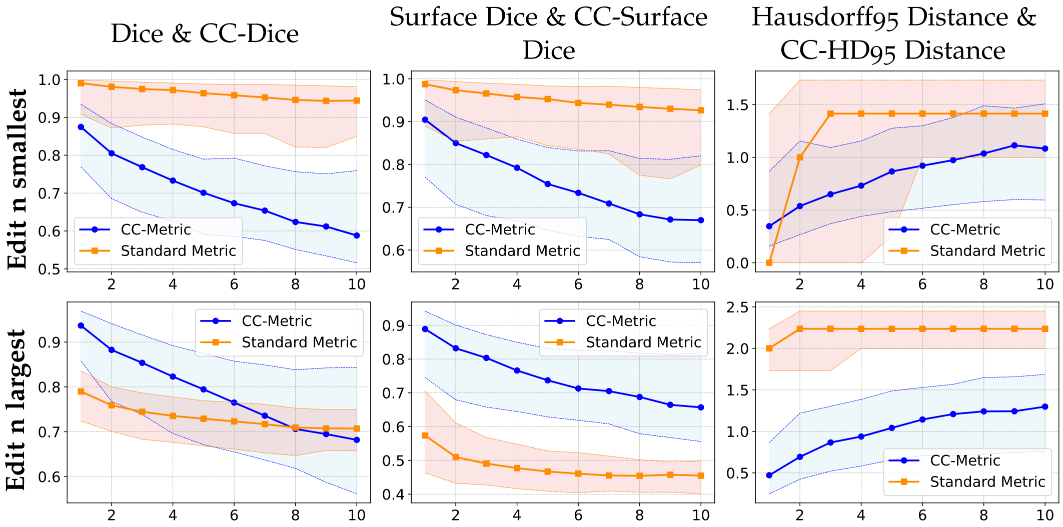


Figure 31: Comparison of standard and CC semantic segmentation metrics on the AutoPET dataset: We enlarge n connected components, simulating that lesions were oversegmented by a model. Standard Metrics are shown in orange, CC-Metrics are shown in blue.

skewed towards large metastases. Even in the scenario where the 10 smallest metastases are not captured, the standard metrics are barely affected. In contrast, CC-Dice and CC-Surface Dice more accurately represent the expected degradation of prediction quality and behave similarly regardless of whether False Negatives are larger or smaller components.

Oversegmentation (Figure 31): We observe that overlap-based metrics as well as boundary-based metrics are dominated by large components. The Distance-based metrics HD95 and CC-HD95 have the capability to capture deviations of small and large components. A major downside of the standard Hausdorff metrics is, however, that they stay constant, reporting the global maximum distance for the standard HD and the global 95th percentile for the HD95 metric. Due to this worst-case behavior of the metrics, a single component dominates the score, rendering how well other components are segmented whose scores fall below the global worst-case irrelevant. CC-HD95 on the other hand reports the average per-component worst case and provides a more nuanced signal, allowing each component to influence the score.

Undersegmentation (Figure 33): In this scenario, we simulate under-segmentation where the model misses parts of the tumor or metastasis in Figure 33. We again observe CC-Metrics offering more nuanced insights into the prediction quality. Note that the standard HD95 metric first ignores errors when degrading the smallest errors as they fall below the 95th percentile due to the large number of global boundary pixels. Again CC-HD95 is more informative.

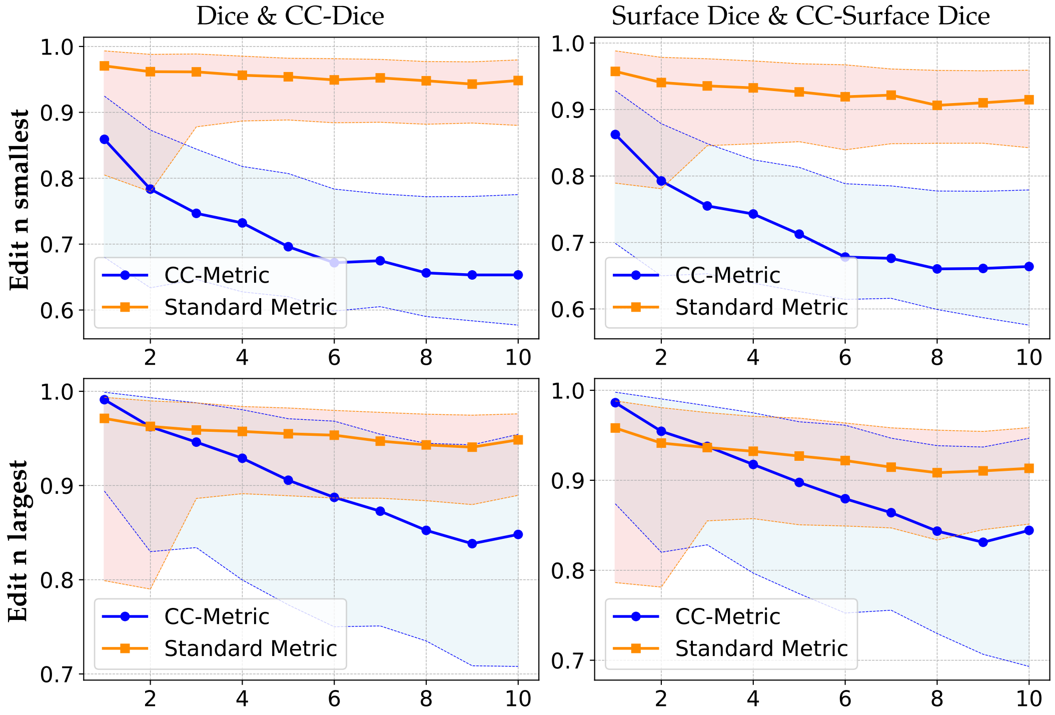


Figure 32: Comparison of standard and CC semantic segmentation metrics on the AutoPET dataset: We insert n random connected components, simulating that lesions were falsely segmented by a model. Standard Metrics are shown in orange, CC-Metrics are shown in blue.

False Positives (Figure 32): In this scenario, we add one component to each of the n selected regions. For each tumor volume, we randomly sample a location within the region defined by the n largest or smallest components and insert a tumor there. The inserted tumors have a volume representing the 25th percentile of all metastasis volumes in the patient. We observe that CC-Dice and CC-Surface Dice are generally more sensitive to false positive predictions than standard Dice and Surface Dice under this simulation scenario.

5.1.5.2 *Evaluation of Real Predictions*

We train 5 segmentation models, namely nnUNet [120], DynUNet [121], UNETR [99], SwinUNETR [98] and MedNext [272]. Except for nnUNet, we use a consistent training process outlined in Appendix E.1.2 across all models with varying capabilities to assess how standard metrics and the novel CC-Metrics evaluate them, focusing on metric comparison rather than optimizing model performance. We report the evaluation of the model predictions in Table 16. For unbound metrics such as the Hausdorff Distance, we set the worst possible score to 50.

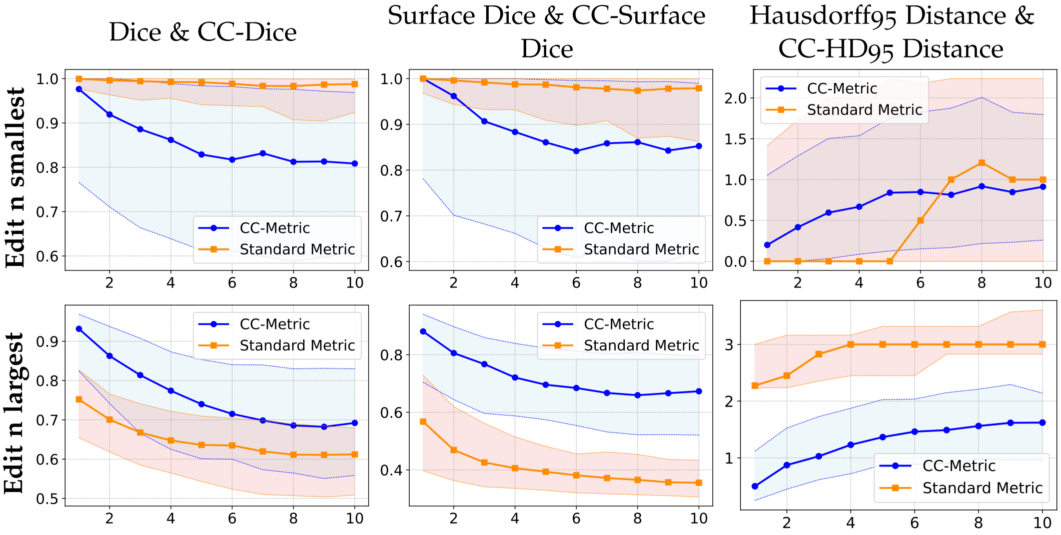


Figure 33: Comparison of standard and CC semantic segmentation metrics on the AutoPET dataset: We perform erosion on n connected components, simulating that lesions were under-segmented by a model. Standard Metrics are shown in orange, CC-Metrics are shown in blue.

On the AutoPET dataset with its high number of average metastases per patient, we find significant differences between standard and CC-Metrics. The CC-Dice scores are significantly lower for all models than the traditional Dice scores. This disparity is most notable in models like nnUNet, where the Dice score is 67.5, but the CC-Dice score drops to 47.4, indicating that while traditional metrics might suggest a model performs well, the large difference between Dice and CC-Dice highlights the model is struggling with smaller metastases. We also find that ranking models by Dice and CC-Dice yields different results, with DynUNet outperforming UNETR in CC-Dice despite its significantly worse performance measured in standard Dice. Regarding the Surface Dice (SD) score, we observe a similar picture as for the Dice. The Hausdorff95 distance, abbreviated as HD in the table, and its CC variant (CC-HD) show marked differences, with CC-HD scores being generally lower, indicating better performance than the standard HD95. This result is to be expected as the standard metric focuses on global worst-case scenarios. The difference can be interpreted as the difference between the global worst-case and the average per-component worst-case scenario for the dataset. This effect is best shown in the simulation results in Figures 31 and 33.

On the HECKTOR dataset, we find CC-Metrics to be very well aligned with standard metrics for all models due to the low average number of ground truth components, highlighting that CC-Metrics do not bias results in unexpected ways.

Table 16: Comparison of standard metrics against CC segmentation metrics

AutoPET: High Metastases Count Dataset

Models	Dice	CC-Dice	SD	CC-SD	HD	CC-HD
nnUNet	67.5	47.4	66.3	52.5	71.1	46.4
DynUnet	62.5	44.0	61.8	49.1	127	22
MedNext	67.7	46.1	66.91	51.62	73.8	28.18
UNETR	41.6	28.0	33.9	29.3	211	157
SUNETR	54.8	38.4	49.0	40.2	174	130

HECKTOR: Low Metastases Count Dataset

Models	Dice	CC-Dice	NSD	CC-SD	HD	CC-HD
DynUnet	78.7	78.7	80.3	80.3	37.7	37.6
UNETR	60.8	60.7	56.8	56.8	151	151
SUNETR	72.3	72.3	71.2	71.3	113	113

5.1.6 *Qualitative Results*

In Figure 34 we show two examples of nnUNet predictions where standard Dice and CC-Dice disagree by a large margin. On the left, the Dice score of the prediction is 46.7%, while CC-Dice was reported as 74.4%. Prior to analyzing the qualitative example, one could hypothesize that the observed discrepancy might stem from the model's tendency to more accurately capture smaller predictions compared to larger ones. This is confirmed when observing the left plot, where the three small metastases are covered by the model, whereas the large metastasis in the right cheek is missed. This is however a rather rare example in the nnUNet predictions, as most of the patient-wise CC-Dice scores are lower than their standard Dice scores. A typical example where CC-Dice is much lower than standard Dice is shown on the right. The reported standard Dice for this example is 85.3%, whereas CC-Dice is only at 20.1%. While the predictions are well aligned, at first sight, there are subtle differences. For instance, a false positive is being segmented close to the aorta and a metastasis which is located on the right of the largest connected component (indicated by the arrow) has not been segmented by the network. Other metastases are either over- or undersegmented. While these subtle differences are not captured by the standard dice, some of these can have detrimental effects on the patient's projected survival time and may require a sudden change in the current cancer treatment plan. This example also gives an intuition on how the difference in standard metrics and CC-Metric reveals the patterns of errors in the network predictions.

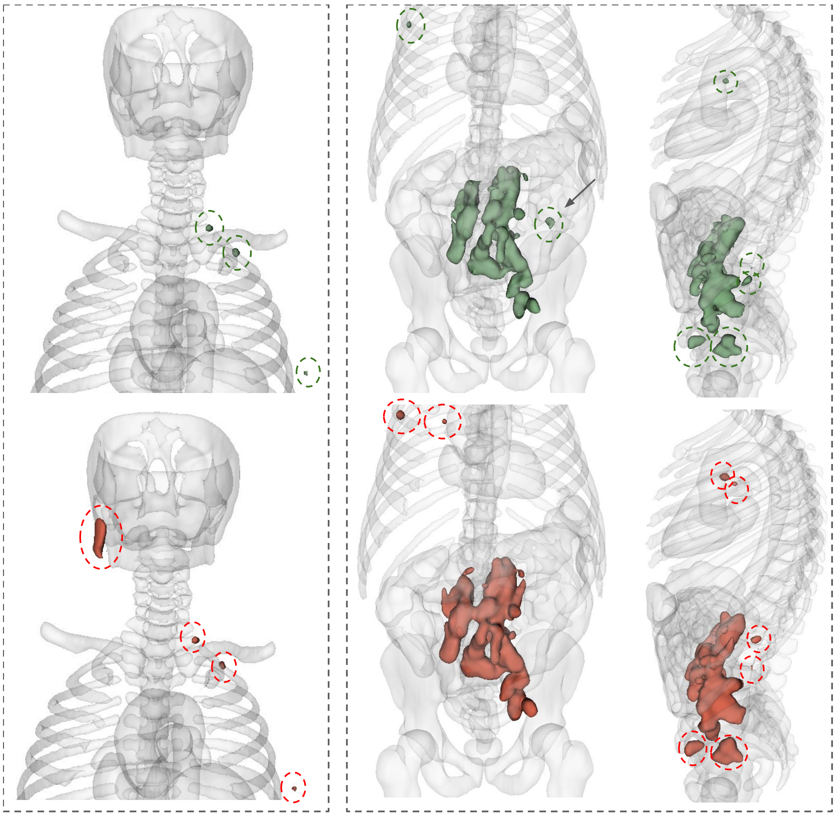


Figure 34: Qualitative Example of predictions (red) vs. ground truth (green) with large differences in standard Dice and CC-Dice. On the left, CC-Dice (74.4%) is higher than the standard Dice (46.7%), while on the right, the standard Dice (85.3%) exceeds CC-Dice (20.1%).

5.2 CHAPTER CONCLUSION

In this chapter, we introduced CC-Metrics, a novel evaluation protocol that addresses critical limitations of traditional semantic segmentation metrics when applied to multi-instance lesion detection scenarios. By leveraging generalized Voronoi diagrams to partition images according to ground truth connected components, CC-Metrics enables the computation of established metrics on a per-component basis, ensuring equal weighting of all lesions regardless of size, thereby aligning closer with the clinical reality for which these models are developed. Beyond its methodological contributions, it is important to clarify the assumptions underlying CC-Metrics and the imaging domains to which it can be reliably applied. The proposed proximity-based segmentation-to-ground-truth matching assumes a constant real-world distance between neighboring pixels, which holds for modalities like MRI and CT scans. Under this assumption, CC-Metrics is also applicable to other medical domains with uniform pixel spacing, such as PET, ultrasound (with isotropic resampling), and digital histopathology, as well as to natural imaging domains like satellite imagery, aerial orthophotos, microscopy, industrial X-ray inspection, and remote sensing hyperspectral data. In contrast, it is not suited for images where perspective induces non-uniform pixel distances, such as standard photography or videos from moving cameras. The equidistance criterion is necessary because Voronoi-based matching assumes that distances between neighboring pixels correspond to equal real-world distances. If this is not the case, as in images with perspective, the matching would become location-dependent and could produce inconsistent results.

Our comprehensive analysis demonstrates that standard metrics such as Dice coefficient, Hausdorff distance, and Surface Dice exhibit significant bias toward larger components, potentially masking poor performance on smaller but clinically critical lesions. While existing unified segmentation and detection metrics like Panoptic Quality and Lesion Dice attempt to combine segmentation and detection performance, they suffer from threshold dependencies, non-unique matching, and inconsistent behavior across different scenarios. In contrast, CC-Metrics provides the most general solution by naturally extending any existing segmentation metric—including overlap-based, boundary-based, and distance-based measures—to the per-component evaluation paradigm without introducing additional hyperparameters or matching complexities. Through extensive validation on synthetic predictions and real model evaluations across AutoPET and HECKTOR datasets, we show that CC-Metrics provides more nuanced and clinically relevant assessments of model performance, revealing failure modes that traditional metrics fail to capture. The proposed framework is hyperparameter-free, computationally efficient, and maintains consistency with existing metrics in low-complexity scenarios while providing superior discriminative power in multi-instance settings. These findings have important implications for the clinical translation of lesion segmentation models, as CC-Metrics can help identify models that reliably detect both large and small metastases—a capability essential for accurate staging and treatment planning in cancer care. The adoption of CC-Metrics in the evaluation pipeline represents a significant step toward developing more robust and clinically reliable segmentation models for medical imaging applications.

Contribution 1: We demonstrate critical limitations of standard semantic segmentation metrics when applied to multi-instance lesion detection, showing through systematic analysis that overlap-based, boundary-based, and distance-based metrics exhibit significant bias toward larger components, potentially masking poor performance on clinically critical smaller lesions.

Contribution 2: We introduce CC-Metrics, a novel evaluation protocol based on generalized Voronoi diagrams that enables per-component evaluation of any existing segmentation metric. Unlike existing unified metrics such as Panoptic Quality and Lesion Dice, our approach is hyperparameter-free, avoids threshold dependencies, and naturally extends to all metric types without introducing matching complexities.

Contribution 3: We extensively validate CC-Metrics on synthetic and real network predictions, demonstrating that our per-component evaluation approach reveals clinically relevant failure modes invisible to standard metrics. We find substantial disagreements between CC-Metrics and traditional evaluation protocols, underscoring the importance of selecting proper metrics to translate technical progress into clinical outcomes.

Part III

INSIGHTS AND FUTURE WORK

6 IMPACT ON THE FIELD

This thesis advances medical image segmentation, with a particular focus on radiological imaging and the integration of anatomical knowledge throughout the model development pipeline. In this chapter, we synthesize the key contributions and outline how this work has driven progress across multiple dimensions: new research directions, novel datasets, innovative methods and models, and advanced evaluation protocols.

6.1 NEW RESEARCH DIRECTIONS

6.1.1 *The Role of Label Quality for Pretraining*

Segmentation algorithms rely on annotated datasets, but our analysis of recent large-scale datasets reveals significant labeling errors. We expand existing research on noisy labels in Section 3.2 by examining volumetric segmentation across multiple CT datasets and incorporating multi-label data. Using foundation model-generated pseudolabels, we simulate realistic label construction methods and present the first studies on the influence of label quality in pretraining scenarios of medical imaging. Our findings hint towards the irrelevance of smaller inaccuracies under a sufficiently large dataset for in-domain evaluation, but more importantly, we find that label quality may have minimal effects on pretraining outcomes, hinting at the possibility that improving label quality might not be as valuable if datasets are used primarily for pretraining.

We offer a new perspective on label quality in pretraining for segmentation algorithms. Our findings indicate that label quality had only minimal effects, suggesting that higher label quality in pretraining datasets does not necessarily translate into improved model performance. This could have significant implications for resource allocation in dataset curation and model training. Additionally, our research highlights the potential of foundation model-generated pseudolabels as an alternative to manual annotations, supporting our DAP-Atlas generation process.

6.1.2 *Non-Physical Interactions for Interactive Models*

In Section 3.3, we introduce LIMIS, the first language-based interactive segmentation system that could be used without physical input devices. Leveraging this setup allows segmentation refinement during active medical procedures where hands are occupied, such as orthopedic surgeries, endoscopic procedures, and cardiac catheterization, thereby expanding interactive segmentation from pre- or post-procedure analysis to real-time clinical applications.

We demonstrate how natural language can serve as an alternative to traditional click-based interactions in image segmentation by translating physical actions into constrained, language-operable settings and adapting active learning procedures to enable a dialogue between human and system. For common errors, LIMIS provides guided approaches for issues like region misidentification and over/under-segmentation, potentially improving segmentation outcomes across different user experience levels.

LIMIS provides evidence that sophisticated image analysis tasks can maintain precision while becoming more accessible through conversational interfaces. This opens possibilities for enhanced accessibility, remote interaction capabilities, and integration with voice recognition systems, potentially informing the design of future medical AI tools that prioritize natural interaction paradigms.

6.2 NEW DATASET: DAP ATLAS

A central contribution of this work is the DAP-Atlas dataset as introduced in Section 3.1. DAP-Atlas represents the first comprehensive full-body CT dataset, labeling the majority of human body voxels across 142 distinct classes. Due to its innovative construction methodology, the dataset eliminates the need for annotations by medical professionals while still receiving commendable evaluations from radiologists. Additionally, DAP-Atlas encompasses registered FDG-PET scans with expert-annotated tumors, as it builds on top of the AutoPET [81] dataset. This dual-modality framework, combined with integrated anatomical and pathological labels, underscores the dataset's unique value.

6.3 NEW METHODS

6.3.1 *Joint Anatomy-Pathology modelling*

In Section 4.1, we present APEx, a dual-decoder architecture that explicitly models anatomy and pathology in parallel, thereby enabling structured information exchange between the two. This design mirrors the diagnostic reasoning of radiologists, whose anatomical expertise sharpens pathology detection. By systematically ablating integration strategies, we demonstrate that shared embeddings, combined with targeted cross-decoder communication, yield consistent gains across modalities and tasks. The impact lies in establishing anatomy-guided segmentation as a general, transferable principle providing a blueprint for future architectures that embed anatomical priors to achieve more robust and clinically aligned pathology predictions. Our approach marks a shift from prior, narrow-focused strategies (often confined to a single organ by explicitly modelling the organ's properties) towards a flexible, anatomical label-based, whole-body paradigm.

6.3.2 *Dual-Anatomy Injection Strategy*

We introduce the GRASP framework in Section 4.2, which establishes a scalable, architecture-agnostic paradigm for anatomy-guided pathology segmentation by reusing existing anatomy segmentation models as frozen knowledge sources. Unlike prior approaches that require paired anatomy–pathology datasets, auxiliary losses, or architectural redesigns, GRASP injects anatomical priors via lightweight feature fusion and pseudo-label inputs, eliminating the need to relearn anatomy from scratch. This decoupling of anatomy acquisition from pathology training enables flexible integration across imaging modalities, tasks, and backbone designs, while preserving inference efficiency. The framework shifts the field from tightly coupled, supervision-heavy anatomy–pathology training toward a more modular, plug-and-play approach to anatomy integration.

6.4 NEW EVALUATION PARADIGM: CC-METRICS

Chapter 5 addresses a long-overlooked gap in the evaluation of lesion segmentation models: the disconnect between standard semantic segmentation metrics and the clinical reality of multi-lesion scenarios. While conventional metrics like Dice, Surface Dice, or Hausdorff95 Distance implicitly overweight large lesions, they systematically underrepresent performance on small but clinically decisive findings. CC-Metrics resolves this bias by applying existing, well-established metrics on a per-component basis, ensuring that each lesion, large or small, contributes equally to the evaluation. Importantly, CC-Metrics does not introduce any new metrics, making them simple to understand and straightforward to adopt; the framework also enables distance-based metrics such as Hausdorff Distance to be applied naturally to multi-instance scenarios, thereby providing more informative average per-component agreements compared to the global worst case.

The key impact is twofold. First, CC-Metrics makes evaluation outcomes more clinically aligned, allowing model developers to detect error patterns such as consistent misses of small lesions that would otherwise be masked by volume-dominated scores. Second, to our knowledge, this is the first approach in medical image segmentation that employs spatial proximity via improved Voronoi-based matching, rather than overlap thresholds, to align predictions with ground-truth instances, enabling consistent and threshold-free multi-instance evaluation. Originally introduced for lesion segmentation in PET/CT, the approach generalizes naturally to any domain with uniform pixel spacing, including MRI, PET, ultrasound, and histopathology, as well as non-medical settings such as microscopy, satellite imagery, and aerial orthophotos. CC-Metrics is most impactful when instance sizes vary substantially, as standard semantic segmentation metrics tend to be dominated by large structures, potentially obscuring model errors on smaller but equally important components.

We hope that this will spark interest in more clinically relevant evaluation practices that better capture the true impact of model errors on patient care.

7 FUTURE WORK

We have worked towards showcasing the role of anatomical knowledge through the model development lifecycle: dataset creation, model training, and model evaluation. Within this chapter, we point towards how our work could be advanced in the future. We develop a case for future research for each of the three model lifecycle steps and briefly outline promising directions that could extend the presented contributions.

7.1 AUTOMATED DATASET GENERATION

With the rise of larger and larger anatomical datasets [130, 177, 178, 256, 333], we discussed the challenges of noisy labels extensively in Sections 3.1 and 3.2. While our analysis points towards the irrelevance of smaller inaccuracies under a sufficiently large dataset size, we find that in pretraining scenarios, neither dataset size nor label quality appears to be critical: smaller datasets can outperform larger ones, and pretraining on low-quality labels yields performance comparable to high-quality labels. However, all pretraining configurations, regardless of size or label quality outperform training from scratch.

These findings raise key questions: What transferable representations are learned during pretraining, and what defines a "good" pretraining dataset in the medical domain? For those using datasets: Should robust loss functions be the default, or is label quality generally sufficient for standard training? For those creating datasets: Is it more effective to remove the worst examples or to uniformly improve all samples? Ultimately, how should limited expert annotation resources be allocated to maximize in-domain and downstream model performance?

7.2 LEVERAGING ANATOMICAL KNOWLEDGE FOR PATHOLOGY SEGMENTATION

In Chapter 4, we presented two cases in which anatomical knowledge, expressed through anatomical labels, supports the segmentation of pathological structures. This formulation follows the premise that a model capable of segmenting anatomy inherently possesses anatomical knowledge. However, this approach is limited to the imaging domain and disregards other valuable information sources, such as patient history in report form or standardized medical knowledge from textbooks, which could enable richer interpretation and contextualization of findings.

A multimodal system combining such external knowledge with learned representations from imaging data could distinguish between universally known medical facts and case-specific evidence, integrate complex reasoning, and generate multimodal outputs. This would enable more sophisticated interactions with clinicians and potentially patients.

With the rapid progress in multimodal vision–language models [90, 174, 386] and agentic workflows [112, 232] that integrate multiple information sources, extending our work toward such systems represents a promising future research direction.

7.3 EVALUATING LESION SEGMENTATION MODELS

While CC-Metrics provides a clinically intuitive and generalizable framework for multi-instance segmentation evaluation, several directions could further increase its impact and adoption.

User studies should assess the acceptance of CC-Metrics among clinicians and researchers, for example, by ranking model outputs and comparing CC-Metrics rankings to expert preferences. Such studies could also identify scenarios where CC-Metrics diverges from human judgment, thereby guiding metric refinements.

Anatomy-based weighting schemes could prioritize instances according to their clinical relevance, such as weighting lesions by anatomical location or suspected metastatic pathway. This could allow CC-Metrics to more accurately reflect tumor spread patterns and to quantify how segmentation errors might alter treatment decisions. In its current implementation, it approximates this behavior by treating all lesions equally.

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

Following the "Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG"[1] from September 2023, the author utilized DeepL and ChatGPT (version 4, 4.5 and 5) to enhance the language quality of this work and to support typesetting tasks (e.g., formatting tables).

[1] https://www.dfg.de/resource/blob/289674/ff57cf46c5ca109cb18533b21fba49bd/230921-stellungnahme-praesidium-ki-ai-data.pdf

Part IV

APPENDIX

ALEXANDER JAUS

PhD Student ◇ Helmholtz Data Science School for Health

Karlsruhe Institute of Technology, Adenauerring 10, 76131 Karlsruhe
alexjaus19@gmail.com ◇ linkedin.com/in/alexander-jaus/ ◇ github.com/alexanderjaus

EDUCATION

Karlsruhe Institute of Technology, Germany *May 2022 - Present*
PhD student in Computer Vision

- Advisor: Prof. Dr. Rainer Stiefelhagen & Prof. Dr. Jens Kleesiek
- Stipend of the HIDSS4Health graduate school formed by KIT, DKFZ, and the University of Heidelberg

Karlsruhe Institute of Technology, Germany *May 2017 - Aug 2021*
M. Sc. Computer Science

- Focus on ML, Statistics, and Deep Learning, Minor: Business (Design Thinking, Finance)
- Design Thinking Innovation Project with KIT, HSG & Stanford University over 9 months
- Thesis in Computer Vision: Generalization of panoptic image segmentation to panoramic images by using a novel continuous contrastive-learning based approach.

Karlsruhe Institute of Technology, Germany *Oct 2012 - Oct 2017*
B. Sc. Industrial Engineering and Management

- Academic Exchange Year, National Taiwan University (NTU), Taipei (2016 - 2017)
- Focus on Statistics, Econometrics, and Machine Learning

WORKING EXPERIENCE

Karlsruhe Institute of Technology, Germany *Oct 2021 - Present*
Researcher at cv:hci Computer Vision Lab

- Published in prestigious conferences like MICCAI or AAAI and high-impact journals such as T-ITS
- Conceptualizing and teaching of a lecture on Diffusion- and other Generative Models
- Teaching supervision for multiple practical courses and seminars on Computer Vision and LLMs
- Technical Infrastructure Administrator: Ensuring the reliability and scalability of a compute cluster for 30+ daily users via LDAP, Kerberos, Nagios, and KVM

SAP Research, France *Oct 2019 - Apr 2020*
Research Intern

- Forecasted time series of system critical KPIs such as RAM or Disk I/O in order to improve the stability of customer systems using Gaussian Process models, CNNs, and LSTMs in TensorFlow
- Benchmarked anomaly detection in forecasted time series, allowing predictive system maintenance

scieneers, Germany *Oct 2019 - Apr 2020*
Data Science Student

- Developed a Python-based digital coaching platform, utilizing biomarkers collected through the wearables of athletes to analyze their performance. Based on collected biomarkers, the coach dynamically updates personalized marathon training plans for optimal progress and goal achievement
- Tested the developed product with experienced athletes over a training period of three months

AWARDS AND HONORS

- Best Paper award for "Panoramic Panoptic Segmentation: Towards Complete Surrounding Understanding via Unsupervised Contrastive Learning" at IEEE Intelligent Vehicles Symposium, 2021
- Second Place in AutoPET III Datacentric Lesion Segmentation challenge held at MICCAI 2024
- First Place at Segment Anything Medical on a Laptop Challenge with Scribbles at CVPR 2024

TEACHING

- **Teaching Assistant**
  - Deep Learning for Computer Vision I: Basics *SS 2023, SS 2024*
  - Deep Learning for Computer Vision II: Advanced Topics *WS 22/23 – WS 25/26*
  - Practical Course: Computer Vision for HCI *SS 2023, SS 2024*
  - Seminar: Computer Vision for HCI *WS 22/23 – WS 24/25*
  - Seminar: Multimodal Large Language Models *SS 2024*
- **Supervision of Master's Theses**
  - Diffusion Models for Medical Image Segmentation
  - Anatomically Plausible Data Augmentation
  - Interactive Segmentation of Medical Images via Language Prompts
  - Leveraging Anatomical Priors for Lesion Segmentation
- **Freelance Lecturer**
  - Software Development using Generative AI Tools at IHK *Sep 2025 - Present*

SERVICES AND OUTREACH

Peer Review

- Journals: T-ITS, TMI
- Conferences: CVPR, ICCV, ECCV, NIPS, BMVC

Other

- Invited Talks: ISBI 2025
- Memberships: AAAI

B

AUTHORED PUBLICATIONS

The work presented in this doctoral thesis resulted in the following thesis-related publications, arranged by publication date. An asterisk (*) indicates equal first authorship by A. Jaus.

[1] Alexander Jaus*, Constantin Seibold*, Kelsey Hermann, Alexandra Walter, Kristina Giske, Johannes Haubold, Jens Kleesiek, and Rainer Stiefelhagen. "Towards unifying anatomy segmentation: automated generation of a full-body ct dataset via knowledge aggregation and anatomical guidelines." In: *arXiv preprint arXiv:2307.13375* (2023).

[2] Alexander Jaus, Constantin Seibold, Simon Reiß, Lukas Heine, Anton Schily, Moon Kim, Fin Hendrik Bahnsen, Ken Herrmann, Rainer Stiefelhagen, and Jens Kleesiek. "Anatomy-guided Pathology Segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer Nature Switzerland Cham. 2024, pp. 3–13.

[3] Alexander Jaus*, Constantin Seibold*, Kelsey Hermann, Negar Shahamiri, Alexandra Walter, Kristina Giske, Johannes Haubold, Jens Kleesiek, and Rainer Stiefelhagen. "Towards Unifying Anatomy Segmentation: Automated Generation of a Full-Body CT Dataset." In: *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2024, pp. 41–47.

[4] Lena Heinemann*, Alexander Jaus*, Zdravko Marinov, Moon Kim, Maria Francesca Spadea, Jens Kleesiek, and Rainer Stiefelhagen. "LIMIS: Towards Language-Based Interactive Medical Image Segmentation." In: *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2025, pp. 1–5.

[5] Alexander Jaus, Zdravko Marinov, Constantin Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. "Good Enough: Is it Worth Improving your Label Quality?" In: *arXiv preprint arXiv:2505.20928* (2025).

[6] Alexander Jaus, Constantin Marc Seibold, Simon Reiß, Zdravko Marinov, Keyi Li, Zeling Ye, Stefan Krieg, Jens Kleesiek, and Rainer Stiefelhagen. "Every Component Counts: Rethinking the Measure of Success for Medical Semantic Segmentation in Multi-Instance Segmentation Tasks." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 4. 2025, pp. 3904–3912.

[7] Keyi Li*, Alexander Jaus*, Jens Kleesiek, and Rainer Stiefelhagen. "GRASPing Anatomy to Improve Pathology Segmentation." In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2025.

The following publications were co-authored by A. Jaus but fall outside the scope of this thesis and are listed in chronological order.

[1] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. "Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning." In: *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. 2021, pp. 1421–1427.

[2] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. "Panoramic panoptic segmentation: Insights into surrounding parsing for mobile agents via unsupervised contrastive learning." In: *IEEE Transactions on Intelligent Transportation Systems* 24.4 (2023), pp. 4438–4453.

[3] Constantin Seibold, Alexander Jaus, Matthias A Fink, Moon Kim, Simon Reiß, Ken Herrmann, Jens Kleesiek, and Rainer Stiefelhagen. "Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling." In: *arXiv preprint arXiv:2306.03934* (2023).

[4] Alexander Jaus, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. "Data Diet: Can Trimming PET/CT Datasets Enhance Lesion Segmentation?" In: *arXiv preprint arXiv:2409.13548* (2024).

[5] Valentin Khan-Blouki, Franziska Seiz, Nicolas Walter, Alexander Jaus, Zdravko Marinov, Gijs Luijten, Jan Egger, Constantin Marc Seibold, Dirk Solte, Jens Kleesiek, et al. "FootCapture: Towards an AR-based System for 3D Foot Object Acquisition through Photogrammetry." In: *Medical Imaging with Deep Learning*. 2024.

[6] Zdravko Marinov, Alexander Jaus, Jens Kleesiek, and Rainer Stiefelhagen. "Filters, thresholds, and geodesic distances for scribble-based interactive segmentation of medical images." In: *Medical Image Segmentation Challenge*. Springer Nature Switzerland Cham, 2024, pp. 39–56.

[7] Zdravko Marinov, Alexander Jaus, Jens Kleesiek, and Rainer Stiefelhagen. "Taking a Step Back: Revisiting Classical Approaches for Efficient Interactive Segmentation of Medical Images." In: *Medical Image Segmentation Challenge*. Springer Nature Switzerland Cham, 2024, pp. 101–125.

[8] David Schneider, Simon Reiß, Marco Kugler, Alexander Jaus, Kunyu Peng, Susanne Sutschet, M Saquib Sarfraz, Sven Matthiesen, and Rainer Stiefelhagen. "Muscles in time: Learning to understand human motion in-depth by simulating muscle activations." In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 67251–67281.

[9] Jianning Li, Zongwei Zhou, Jiancheng Yang, Antonio Pepe, Christina Gsaxner, Gijs Luijten, Chongyu Qu, Tiezheng Zhang, Xiaoxi Chen, Wenxuan Li, et al. "Medshapenet–a large-scale dataset of 3d medical shapes for computer vision." In: *Biomedical Engineering/Biomedizinische Technik* 70.1 (2025), pp. 71–90.

[10] Simon Reiß, Zdravko Marinov, Alexander Jaus, Constantin Seibold, M Saquib Sarfraz, Erik Rodner, and Rainer Stiefelhagen. "Is Visual in-Context Learning for Compositional Medical Tasks within Reach?" In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2025.

[11] Cedric Zöllner, Simon Reiß, Alexander Jaus, Amroalalaa Sholi, Ralf Sodian, and Rainer Stiefelhagen. "Semantic Segmentation for Preoperative Planning in Transcatheter Aortic Valve Replacement." In: *International Workshop on Statistical Atlases and Computational Models of the Heart*. Springer. 2025.

C

ADDITIONAL DETAILS: AUTOMATED DATASET GENERATION

C.1 AUTOMATED DATASET CONSTRUCTION: DAP ATLAS

C.1.1 *Additional Details for the Section on DAP Atlas*

- **Pediatric [139]:** This dataset consists of 359 chest-abdomen-pelvis and abdomen-pelvis CT images of patients between the age of 5 and 16 years. It provides 29 anatomical structures annotated by experts. Patients were selected based on random clinical indications from the university clinic of Children's Wisconsin.

- **Total Segmentator [335]:** The TotalSegmentator dataset is large and diverse with 1024 CT images of different body parts with labels for 104 anatomical structures. The dataset was collected by randomly sampling from the PACs systems of multiple sites. Its annotation is based on an interactive semi-automatic approach. Here, models are first trained on a few manual annotations. These models infer predictions on unlabeled scans which are lastly refined by an expert. This cycle repeats with an ever-increasing number of training images.

- **SegThor [167]:** A dataset consisting of 60 thoracic CTs collected at the Henri Becquerel Center. The patients were selected based on lung cancer or Hodgkin's lymphoma diagnosis. The CTs contain annotations for four organs at risk whose tissues must remain intact during radiation therapy. The annotations of the dataset are provided by an experienced radiotherapist.

- **CT50Abdomen [206]:** The dataset is part of the CT1k Abdomen datasets extension, in which the authors provide 50 abdominal CT images with previously less annotated structures, such as the adrenal glands. Annotations are provided by multiple junior annotators and checked by senior radiologists.

- **MAL Cervix [168]:** This dataset is part of the Beyond the Cranial Vault challenge. It consists of 30 training and 20 testing abdominal CT images acquired via a full bladder drinking protocol and annotated by a trained radiation oncologist. It focuses on the digestive and reproductive systems of female cervical cancer patients.

- **Amos [133]:** A diverse dataset with 500 CT images collected from different scanners and sites covering 15 abdominal organ categories. The selection of patients relates to abdominal tumors or abnormalities examinations. Annotations rely on a combination of junior and senior radiologist labor.

- **RibSeg [355]:** The RibSeg dataset consists of 490 CT Scans taken from publicly available RibFrac [138] dataset. The authors use a semi-automatic morphology-based segmentation approach based on thresholding, point cloud segmentation, and morphological operations. They check the proposed segmentations by hand and refine them if necessary.

- **Verse [282]:** A large dataset for vertebra segmentation. It consists of two subsets and has a total of 374 CT scans of 355 patients from multiple detectors and sites with voxel-wise annotations for individual vertebras. Segmentations have been performed semi-automatically, with initial proposals being generated by an in-house pipeline. The proposals are refined by a team of trained medical students and experts and finally approved by a radiologist with more than 30 years of experience.

- **ATM [367]:** This dataset establishes a benchmark for Airway Tree Modelling by providing 500 chest CT scans from different sites and includes scans of healthy patients, patients with pulmonary diseases, and even noisy COVID-19 CTs. Annotations of the pulmonary airways were performed by a team of three experts, with each radiologist having more than five years of experience.

- **PARSE [201]:** The PARSE dataset is part of the Pulmonary Artery Segmentation Challenge and contains a total of 203 CT images from 203 patients who have been diagnosed with pulmonary nodular diseases. The CTs were generated using devices from two different manufacturers, with data collected from four distinct sites. Each of the images has been annotated by five experts, with each expert having at least five years of experience in the field.

- **Pelvic CT [187]:** A large-scale dataset that focuses on the segmentation of pelvic bone structures such as hip bones or sacrum. It consists of 1184 CT images collected from different source datasets, combining images from multiple sites, scanners, and even metal artifacts. The labeling was conducted by a team of junior and senior radiologists.

C.1.2 *Population and Diagnosis Characteristics*

Our DAP Atlas is similar to AutoPET regarding the age and gender distributions as well as pathological findings. We show a descriptive analysis of the dataset regarding the aforementioned dimensions in Figure 35.

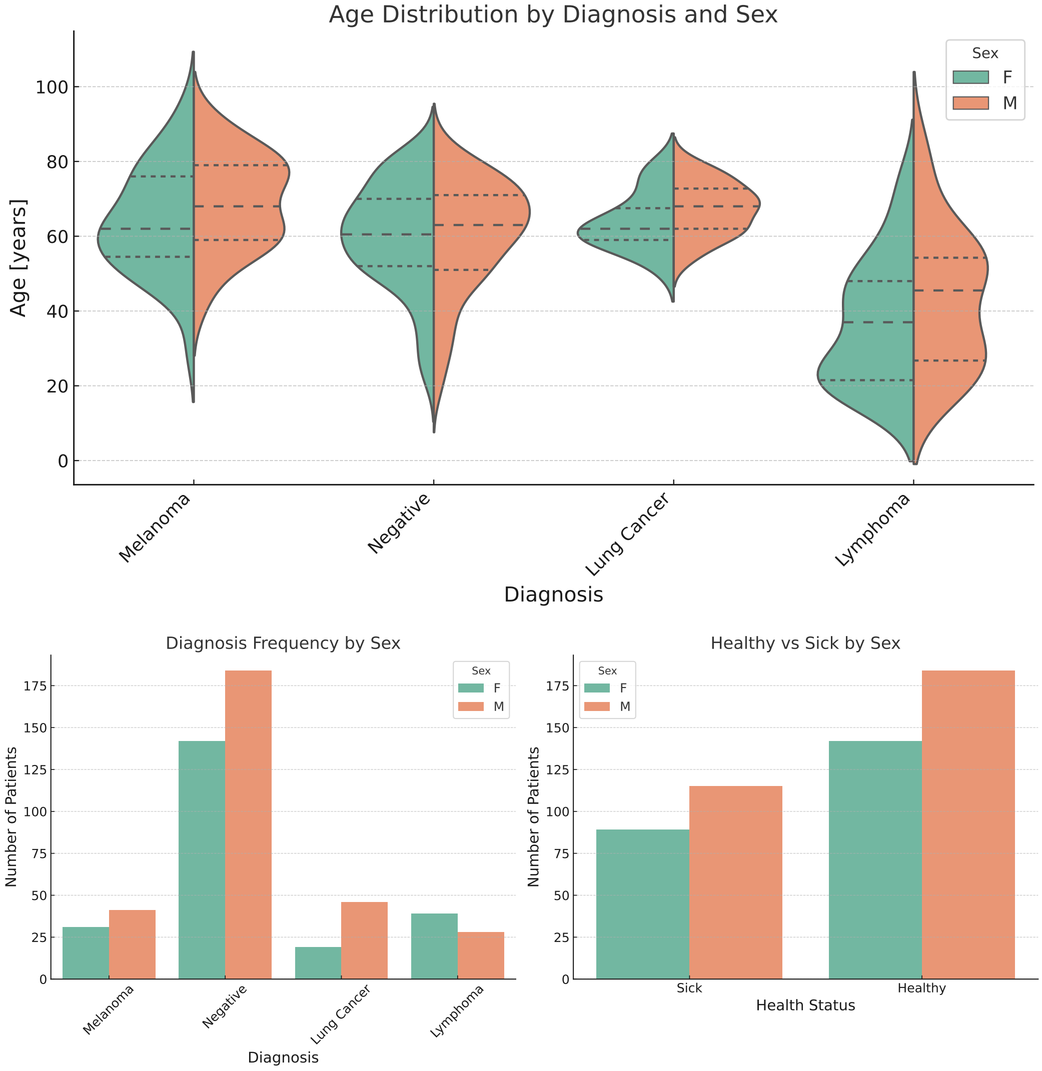


Figure 35: Descriptive statistics of the DAP Atlas dataset. Top: Violin plots display the age distribution stratified by diagnosis and sex. Dashed lines indicate quartiles of the respective distributions. Bottom: Diagnosis frequencies stratified by sex (left panel) and aggregated into healthy vs. sick categories (right panel), illustrating the approximate balance between the two groups.

C.1.3 *Overview of Post-Processing Algorithm*

We show the developed post-processing Algorithm 3. To improve predicted label quality, we use several post-processing approaches combining: Left-Right Split (assigning side-dependent labels correctly), Rib Counting (ordering 24 human ribs), non-largest component suppression (for single-component anatomies like the brain), area restriction (constraining anatomy labels to body parts), and sex-based consistency (ensuring anatomically correct predictions).

C.1.4 *Developing a Prediction Model from the Atlas Dataset*

The goal of the Atlas prediction model is to eliminate the need for post-processing which is impractical within a clinical setting in which the model should be able to deliver convincing results on arbitrary CT volumes. When examining the different steps of Algorithm 3, we notice two steps that are easy to address algorithmically: sex-based consistency and non-largest connected component suppression as defined in Algorithm 4 and Algorithm 5 respectively, as these methods simply suppress predictions and do not rely on other anchor predictions such as Algorithm 8. We thus aim to develop a training procedure that eliminates the need for left-right splitting (Algorithm 8), area-restrictions (Algorithm 6), and rib counting (Algorithm 7).

To tackle these challenges, we develop a custom training strategy for the Atlas prediction model. First, we apply Algorithm 3 during the aggregation phase of the individual expert models to maximize the agreement with the desired output which has been approved by experts. Next, we observe that due to the large number of classes, the standard nnU-Net [120] learning rate schedule is suboptimal as it closely follows a linear learning rate schedule allocating approximately the same number of epochs for small and large learning rates. We find that the proposed task is more difficult than most standard segmentation tasks and thus increase the number of training epochs from 1000 to 5000. Finally, we fine-tune the network for another 1000 epochs with a fixed learning rate of 0.001 and without the standard mirror augmentation. This allows the network to focus on the improvements on smaller structures and helps to mitigate the right-left and rib confusion. We show a comparison of the raw output of the Atlas dataset model, the post-processed volume, and the raw output of the Atlas prediction model in Figure 37. As it can be seen, the output of the robust model has a large agreement with the post-processed predictions of the first model without relying on Algorithm 3. We analyze this behavior and find that the vast majority of predicted structures have an agreement of more than 90% IoU between the post-processed V1 Model and the raw V2 predictions.

Besides the DAP Atlas dataset, we also release the robust segmentation model which can be used to perform inference without post-processing. It furthermore tends to per-

```
Algorithm 3 Post-Processing
Require: Volumetric Model Predictions
Ensure: Refined Volumetric Pseudo-Labels
    Left-Right Split
    Rib Counting
    Non-Largest CC Suppression
    Area Restriction
    Sex-based Consistency
```

```
Algorithm 4 Sex-based Consistency
Require: Repro. Anatomy Predictions; Metadata
Ensure: Sex-Restricted Anatomy Predictions
    Suppress male reproductive anatomies for F patients
    Suppress female reproductive anatomies for M patients
```

```
Algorithm 5 Non-largest CC Suppression
Require: Anatomy occurring once
Ensure: Largest CC for each anatomy
    Identify 3D-CCs for anatomy
    Count voxels of each CC
    Remove non-largest CC
```

```
Algorithm 6 Area Restriction
Require: Predictions; BP associations; anchors
Ensure: Anatomy constrained by BP
    Define BP based on box around anchors
    Bind predictions to associated BP
```

```
Algorithm 7 Rib Counting
Require: Rib Predictions
Ensure: 24 largest CC sorted by height
    Merge rib predictions
    Apply Left-Right Split
    Extract 24 largest 3D-CC
    Order CC by height of median points
```

```
Algorithm 8 Left-Right Split
Require: Side-related labels; Sternum; Vertebrae
Ensure: Side-related labels
    Fit hyperplane through V and S centers
    Remap voxels by center position to hyperplane
```

Figure 36: Overview of the general post-processing algorithm consisting of several sub-procedures to enhance the quality of the pseudolabels developed in Section 3.1

form better for out-of-distribution tasks, which are common within a clinical setting. This behavior can be seen in Table 3 and on a qualitative example in Figure 37.

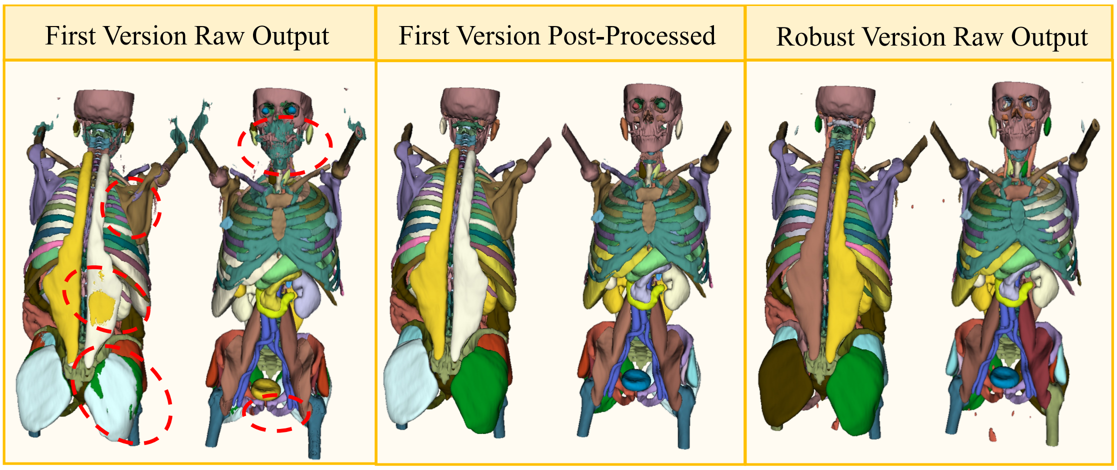


Figure 37: Comparison of the raw output of the standard unified DAP Atlas dataset model, the post-processed volume using Algorithm 3 and the raw output of the more robust Atlas Prediction model. In red we mark problematic regions in the raw labels obtained by the first version of the model. The post-processed volumes and the raw model volumes are alike.

C.2 ADDITIONAL DETAILS FOR THE SECTION ON THE RELEVANCE OF LABEL QUALITY

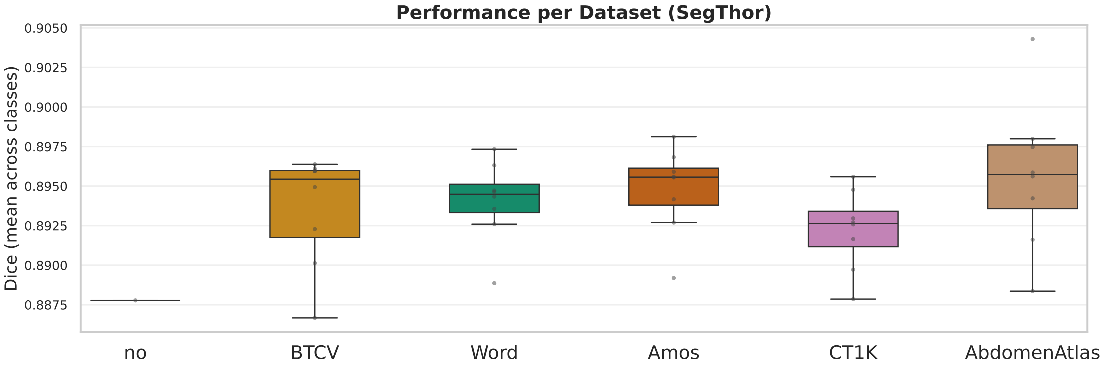


Figure 38: Boxplot comparing the effect of pretraining on different variants of base datasets across different pseudolabel generators, including original data. The models were pretrained on the respective source dataset variants as generated by the pseudolabel generators and subsequently fine-tuned on the clean SegThor [167] dataset, on whose test set the respective model performance was evaluated.

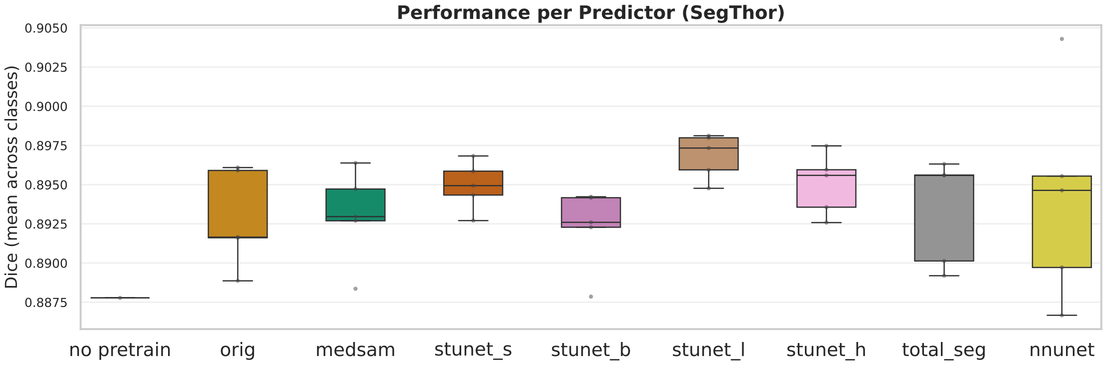


Figure 39: Boxplot comparing the effect of pretraining across datasets for different pseudolabel generators. The models were pretrained on the respective source dataset variants as generated by the pseudolabel generators and subsequently fine-tuned on the clean SegThor [167] dataset. The results are aggregated by comparing the influence of the pseudolabel generator.

D

ADDITIONAL DETAILS: ANATOMICAL KNOWLEDGE FOR PATHOLOGY MODEL TRAINING

D.1 ADDITIONAL DETAILS FOR SECTION ON APEX

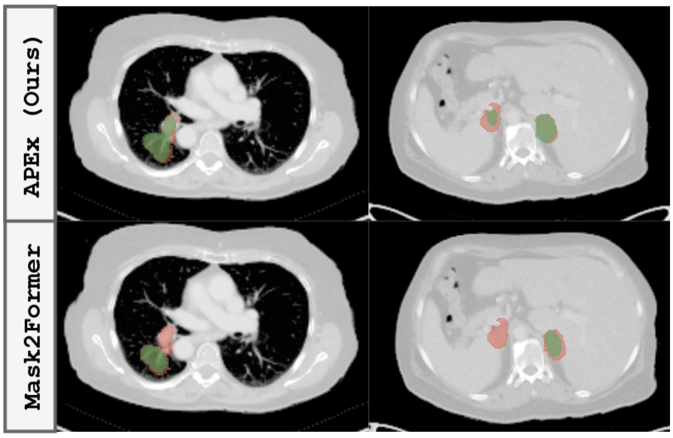


Figure 40: We show qualitative comparisons of APEx against a Mask2Former baseline. The red ground truth is compared against model predictions in green. APEx generally produces more precise structural delineations compared to the baseline approach. We discuss this property in Section 4.1.2.1.

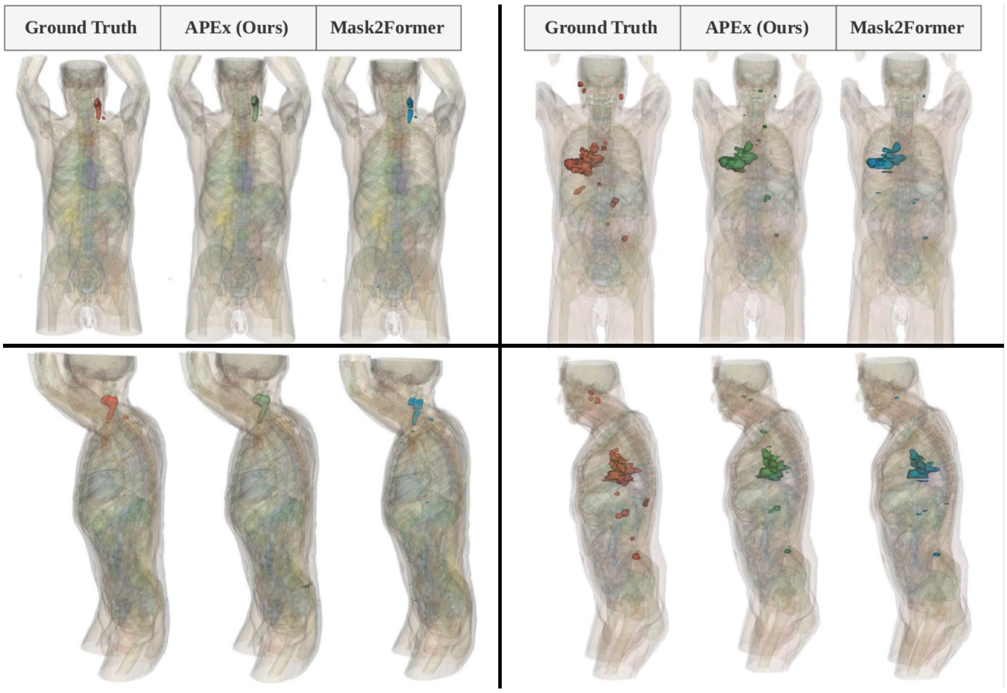


Figure 41: We show qualitative comparisons of APEx against a Mask2Former baseline for two patients in a coronal (top) and sagittal view (bottom). Volumes shown in red, green, and blue denote the lesion ground truth, APE predictions, and Mask2Former baseline predictions, respectively. As discussed in Section 4.1.2.1, APEx, despite being trained exclusively on 2D slices, aligns much closer to the ground truth compared to its Mask2Former baseline.

E ADDITIONAL DETAILS: EVALUATING LESION SEGMENTATION MODELS

E.1 ADDITIONAL DETAILS FOR SECTION ON CC-METRICS

E.1.1 *Details on Synthetic Predictions and additional Simulation Results*

All synthetic predictions are simulated and evaluated on a Red Hat Linux machine with 152 cores and 256GB of RAM. To compute the used semantic segmentation metrics, their respective MONAI [38] implementations are used. We compute CC-Metrics using Algorithm 1 and run the same MONAI implementations per Voronoi Region R'_C.

E.1.1.1 *Evaluation Results on a fixed data Subset*

Figures 30 to 33 in Chapter 5 present the evaluation results of synthetic predictions, where a degrading function progressively worsens an initially perfect prediction over n editing steps. At each step, the analysis includes the maximum available data points, such as all patients with at least n metastases when n metastases are being edited. In the scenario where components are dropped, at least $n+1$ metastases are required when dropping n. While this approach maximizes the number of patients considered at each step, it also involves a different subset of patients at each step. Here, we present a complementary set of Figures 42 to 45, where the subset of patients is kept constant by limiting the analysis to those with at least 10 metastases. For the "Drop n Components" scenario, we accordingly limit the analysis to patients with at least 11 metastases. The observations reported in the main paper remain valid in this scenario with a constant patient subset and are even more pronounced.

E.1.2 *Benchmark Model Training*

E.1.2.1 *Datasets*

We base our experiments on the publicly available AutoPET-II [81] and HECKTOR [241] datasets to ensure a comprehensive analysis across different cancer types. The AutoPET-II dataset, which we also refer to as AutoPET for convenience, includes 1014 samples, consisting of patients diagnosed with malignant melanoma, lymphoma, or lung cancer, as well as negative control patients, whereas the HECKTOR dataset

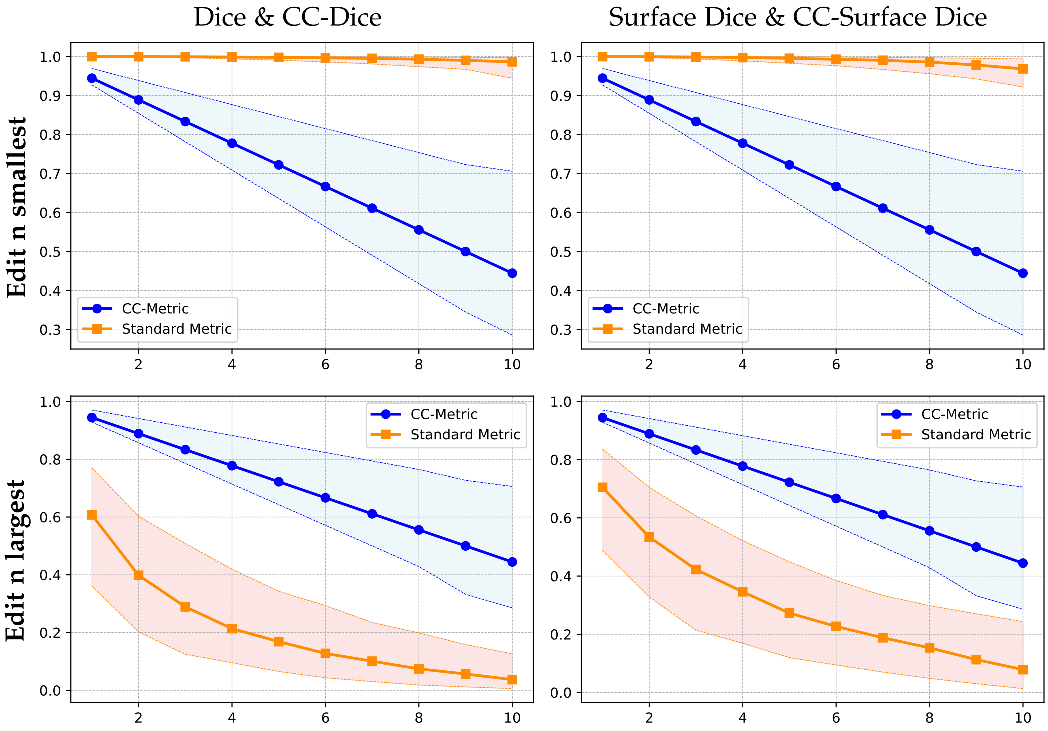


Figure 42: Comparison of standard and CC semantic segmentation metrics on the AutoPET dataset: We drop n connected components, simulating that lesions were forgotten by a model. In this setting, we want to evaluate CC-Metrics on a constant subset of patients. Thus to drop a maximum of $n = 10$ components, we include only patients with at least 11 metastases for all measurements in this scenario including the ones where we drop less than 10 components.

specifically targets head and neck tumors. Another distinguishing characteristic is the average number of tumors per patient in the two datasets: AutoPET features an average of more than 10 tumors per patient, in contrast to HECKTOR, which has only two tumors per patient on average. By utilizing both datasets, our experiments benefit from a diverse range of PET/CT data, providing a realistic assessment of the proposed CC-Metrics across scenarios with both high and low tumor counts.

E.1.2.2 *Data Preprocessing*

As a first step, we align the CT and PET modalities in the HECKTOR dataset. To achieve this, we resample the PET to the CT resolution using third-order spline interpolation. This step is not necessary for AutoPET, as the authors release aligned images. As the normalization procedure, we take inspiration from the nnUNet [120] framework and compute the fingerprints of the two datasets by computing the mean and the standard deviation of the CT and PET image values of the foreground regions. We

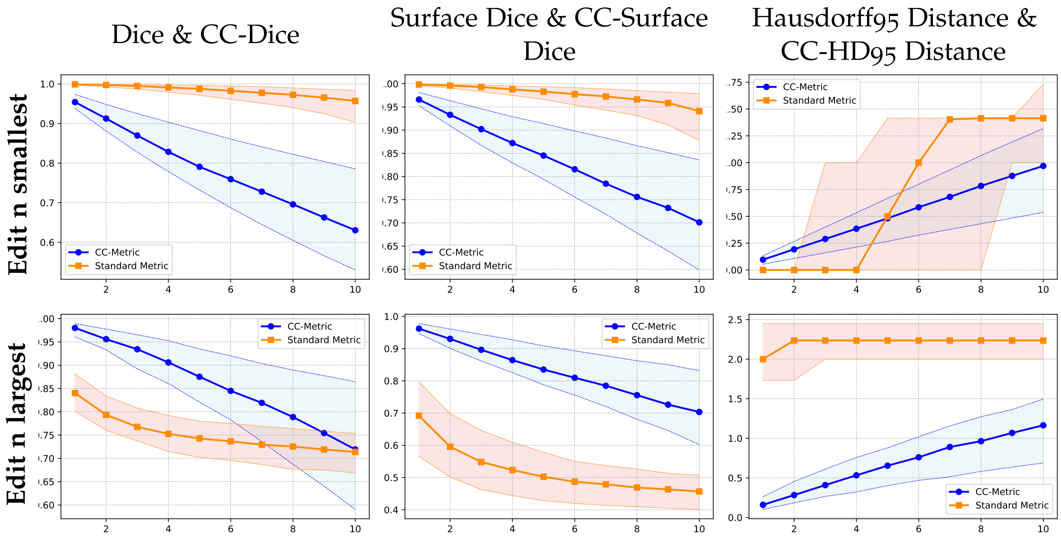


Figure 43: Comparison of standard and CC semantic segmentation metrics on the AutoPET dataset: We enlarge n connected components, simulating that lesions were oversegmented by a model. In this setting, we report on a fixed base of patients and thus consider only the subset of scans that display patients with at least 10 components.

employ a 5-fold cross-validation approach, where the dataset is divided into 5 folds. In each iteration, we train on 4 folds and predict on the remaining fold. This process is repeated 5 times to generate predictions for the entire dataset. To ensure representative folds, we use stratified sampling to generate the individual folds. For the AutoPET dataset, we maintain consistent distributions of cancer types and gender ratios across all folds. In the HECKTOR dataset, which lacks detailed cancer-type descriptions, we stratify only by gender. For all MONAI models, we exclude healthy patients to eliminate pure background samples, resulting in more informative gradients that accelerate the training process. This procedure is not required for nnU-Net, which is inherently trained for 1000 epochs.

E.1.2.3 *Implementation and Training Details*

We evaluate CC-Metrics along with standard evaluation protocols for four trained networks: nnUNet [120], DynUNet [121], UNETR [99], and SwinUNETR [98]. For the nnUNet, we leverage its out-of-the-box capabilities and do not change the default training setting. To handle both the CT and the PET input of the given PET/CT datasets, we concatenate the CT and the PET image as two channels and leave a third channel empty. For the training of DynUNet, UNETR, and SwinUNETR we use their respective MONAI [38] implementations. We clip the CT image values between the 50th and the 95th percentile and the PET image values between the 1st to 99.9th percentile. We

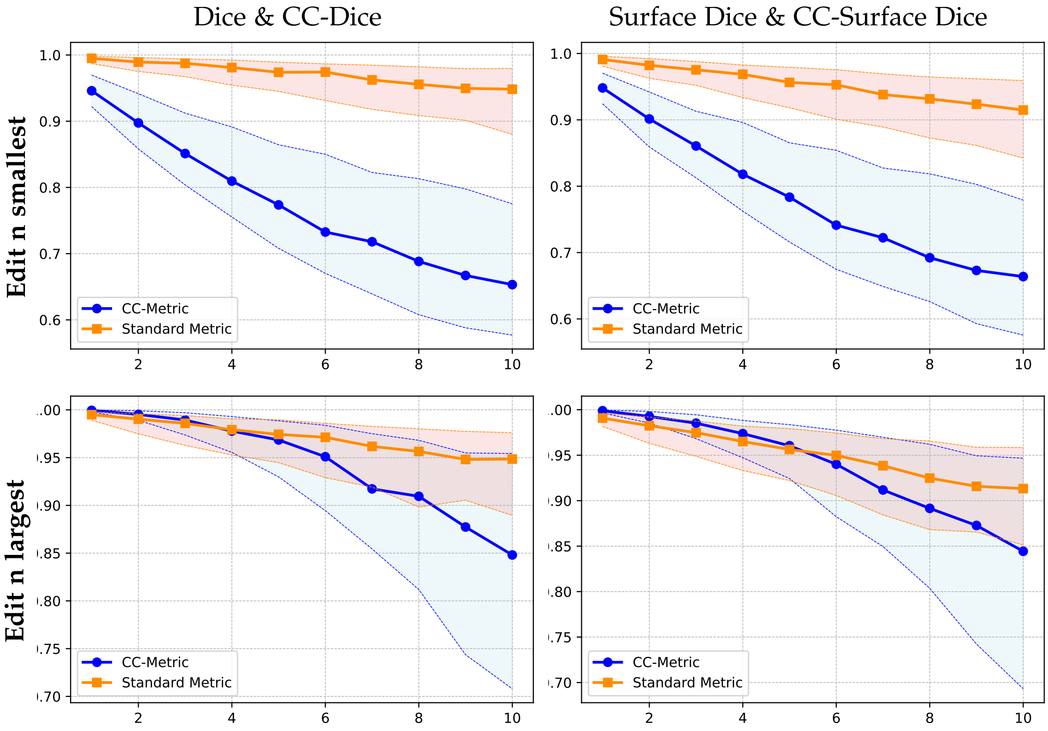


Figure 44: Comparison of standard and CC semantic segmentation metrics on the AutoPET dataset: We insert n random connected components, simulating that lesions were falsely segmented by a model. In this setting, we report on a fixed base of patients and thus consider only the subset of scans that display patients with at least 10 components.

apply a Z-score normalization to each channel by subtracting the mean and dividing by the standard deviation, using the precomputed values from the entire CT and PET datasets.

The training for the Dynamic UNet on AutoPET was conducted over 200 epochs, with validations every 100 epochs. The training process utilized a sliding window approach with a patch size of (128, 128, 128) on AutoPET and (96, 96, 96) on HECKTOR as the images are spatially smaller. We crop 2 patches per image and use a batch size of 8. We use deep supervision and employ the AdamW [199] optimizer to update gradients produced by the DiceCELoss function. We set the initial learning rate to 10^{-3} and gradually reduce it using a Cosine Annealing LR scheduler. For the training of the transformer-based models, we use a patch size of (96, 96, 96) following the settings of the authors [98, 99]. We initialize the SwinUNETR model using the pre-trained weights provided by the authors. The transformer-based models are trained with an initial learning rate of 10^{-4}, which is reduced using a Cosine Annealing scheduler.

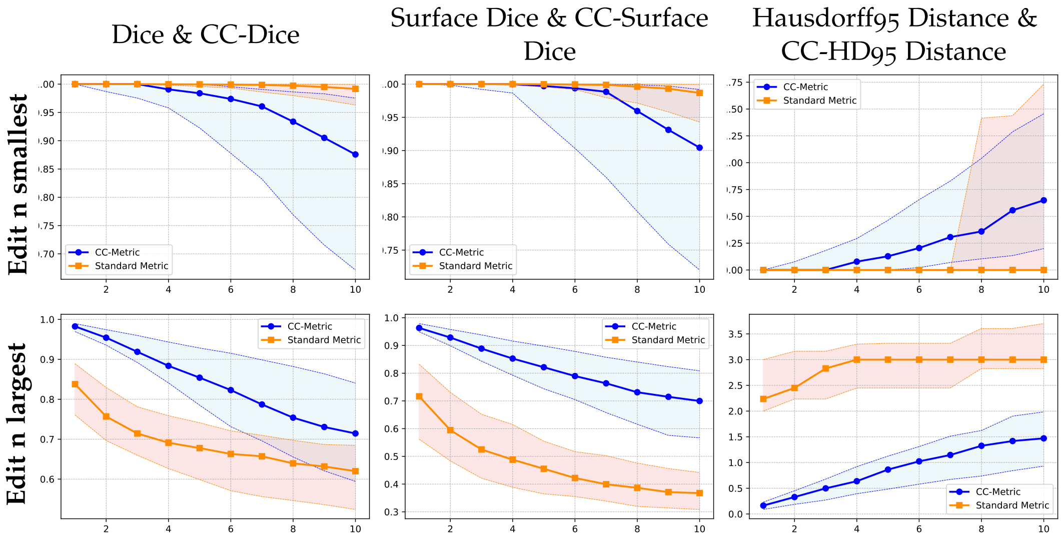


Figure 45: Comparison of standard and CC semantic segmentation metrics on the AutoPET dataset: We shrink n connected components, simulating that lesions were undersegmented by a model. In this setting, we report on a fixed base of patients and thus consider only the subset of scans that display patients with at least 10 components.

Except for the nnUNet, all models are trained with a batch size of 8 on 4 Nvidia A100 GPUs using a DDP-Setting, with 512GB of RAM on a Red Hat Linux machine with 152 cores. We parallelize the training using the PyTorch-Lightning framework [68].

E.1.3 *Proof on the Uniqueness of Generalized Voronoi Regions*

We want to ensure that the same target segmentation mask S always leads to the same Voronoi regions for a given segmentation mask. We verify this for our generalized Voronoi diagram as specified in Definition 4.

$$
\begin{aligned}
R'_C = \{ & t \in \mathbb{T} \mid d'(t, J_{C_k}) < d'(t, J_{C_l}) \\
& \forall C_k, C_l \in \mathbb{K},\ C_k \neq C_l \}
\end{aligned}
$$

Theorem 1. *The generalized Voronoi Diagram as stated in Definition 4 is a unique separation of a metric space.*

Proof: It is a well-known fact that a standard Voronoi diagram is a unique separation of a metric space such as $\mathbb{T}$. We consider the case that in this standard diagram, each site consists of exactly one location (voxel) in each connected component C.
Without the loss of generality, we choose one of the connected components and add one location which is 26-connected to the one that is currently forming the site. This

operation may have the effect that some critical voxel locations $K_{\mathtt{crit}}$ which have previously been within different regions are now closer to the enlarged site. These voxels in $K_{\mathtt{crit}}$, however, would only be closer to the now enlarged site, not an arbitrary other site. This still forms a unique separation of the metric space.

In the context of a discrete metric space, such as $\mathbb{T}$, the situation of boundary points is handled explicitly by the discrete nature of the space. Here, each point (voxel) is either part of a specific region or, in rare cases where distances are exactly equal, part of a shared boundary. In such cases, a consistent rule for boundary assignment is applied, ensuring the uniqueness of the separation. For example, boundary voxels may be assigned to the region associated with the site having the smallest index or based on a deterministic tie-breaking rule.

The described procedure of adding a location to a site can be repeated until all sites reflect $\mathbb{K}$ which fulfills Definition 4 of the Generalized Voronoi region, still forming a unique separation of $\mathbb{T}$. □

BIBLIOGRAPHY

[1] Nebil Achour, Tomas Zapata, Yousef Saleh, Barbara Pierscionek, Natasha Azzopardi-Muscat, David Novillo-Ortiz, Cathal Morgan, and Mafaten Chaouali. "The role of AI in mitigating the impact of radiologist shortages: a systematised review." In: *Health and Technology* (2025), pp. 1–13.

[2] Ibtihaj Ahmad, Sadia Jabbar Anwar, Bagh Hussain, Atiq ur Rehman, and Amine Bermak. "Anatomy guided modality fusion for cancer segmentation in PET CT volumes and images." In: *Scientific Reports* 15.1 (2025), p. 12153.

[3] Mohamed Amgad, Habiba Elfandy, Hagar Hussein, Lamees A Atteya, Mai AT Elsebaie, Lamia S Abo Elnasr, Rokia A Sakr, Hazem SE Salem, Ahmed F Ismail, Anas M Saad, et al. "Structured crowdsourcing enables convolutional segmentation of histology images." In: *Bioinformatics* 35.18 (2019), pp. 3461–3467.

[4] Mario Amrehn, Sven Gaube, Mathias Unberath, Frank Schebesch, Tim Horz, Maddalena Strumia, Stefan Steidl, Markus Kowarschik, and Andreas Maier. "UI-Net: Interactive artificial neural networks for iterative image segmentation based on a user model." In: *arXiv preprint arXiv:1709.03450* (2017).

[5] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. "The medical segmentation decathlon." In: *Nature communications* 13.1 (2022), p. 4128.

[6] Diego Ardila, Atilla P Kiraly, Sujeeth Bharadwaj, Bokyung Choi, Joshua J Reicher, Lily Peng, Daniel Tse, Mozziyar Etemadi, Wenxing Ye, Greg Corrado, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography." In: *Nature medicine* 25.6 (2019), pp. 954–961.

[7] Samuel G Armato III, Karen Drukker, Feng Li, Lubomir Hadjiiski, Georgia D Tourassi, Roger M Engelmann, Maryellen L Giger, George Redmond, Keyvan Farahani, Justin S Kirby, et al. "LUNGx Challenge for computerized lung nodule classification." In: *Journal of Medical Imaging* 3.4 (2016), pp. 044506–044506.

[8] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, Michael F McNitt-Gray, Charles R Meyer, Anthony P Reeves, Binsheng Zhao, Denise R Aberle, Claudia I Henschke, Eric A Hoffman, et al. "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans." In: *Medical physics* 38.2 (2011), pp. 915–931.

[9] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. "A closer look at memorization in deep networks." In: *International conference on machine learning*. PMLR. 2017, pp. 233–242.

[10] Anisha Arulappan and Ajith Bosco Raj Thankaraj. "Liver tumor segmentation using a new asymmetrical dilated convolutional semantic segmentation network in CT images." In: *International Journal of Imaging Systems and Technology* 32.3 (2022), pp. 815–830.

[11] Eirini Arvaniti, Kim S Fricker, Michael Moret, Niels Rupp, Thomas Hermanns, Christian Fankhauser, Norbert Wey, Peter J Wild, Jan H Rueschoff, and Manfred Claassen. "Automated Gleason grading of prostate cancer tissue microarrays via deep learning." In: *Scientific reports* 8.1 (2018), p. 12054.

[12] Murtaza Ashraf, Willmer Rafell Quinones Robles, Mujin Kim, Young Sin Ko, and Mun Yong Yi. "A loss-based patch label denoising method for improving whole-slide image analysis using a convolutional neural network." In: *Scientific reports* 12.1 (2022), p. 1392.

[13] Stefanie Asin. "The Radiologist Shortage, Explained." In: *Becker's Hospital Review* (Dec. 2024). Accessed 2025-06-09.

[14] Marc A Attiyeh, Jayasree Chakraborty, Alexandre Doussot, Liana Langdon-Embry, Shiana Mainarich, Mithat Gönen, Vinod P Balachandran, Michael I D'Angelica, Ronald P DeMatteo, William R Jarnagin, et al. "Survival prediction in pancreatic ductal adenocarcinoma by quantitative computed tomography image analysis." In: *Annals of surgical oncology* 25 (2018), pp. 1034–1042.

[15] Marc A Attiyeh, Jayasree Chakraborty, Lior Gazit, Liana Langdon-Embry, Mithat Gonen, Vinod P Balachandran, Michael I D'Angelica, Ronald P DeMatteo, William R Jarnagin, T Peter Kingham, et al. "Preoperative risk prediction for intraductal papillary mucinous neoplasms by quantitative CT image analysis." In: *Hpb* 21.2 (2019), pp. 212–218.

[16] Alessia Atzeni, Loic Peter, Eleanor Robinson, Emily Blackburn, Juri Althonayan, Daniel C Alexander, and Juan Eugenio Iglesias. "Deep active learning for suggestive segmentation of biomedical image stacks via optimisation of Dice scores and traced boundary length." In: *Medical image analysis* 81 (2022), p. 102549.

[17] Netta Avnoon and Amalya L Oliver. "Nothing new under the sun: Medical professional maintenance in the face of artificial intelligence's disruption." In: *Big Data & Society* 10.2 (2023), p. 20539517231210269.

[18] Wenjia Bai, Ozan Oktay, Matthew Sinclair, Hideaki Suzuki, Martin Rajchl, Giacomo Tarroni, Ben Glocker, Andrew King, Paul M Matthews, and Daniel Rueckert. "Semi-supervised learning for network-based cardiac MR image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 253–260.

[19] Wenjia Bai, Wenzhe Shi, Declan P O'regan, Tong Tong, Haiyan Wang, Shahnaz Jamil-Copley, Nicholas S Peters, and Daniel Rueckert. "A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: application to cardiac MR images." In: *IEEE transactions on medical imaging* 32.7 (2013), pp. 1302–1315.

[20] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian R Rokuss, Ziyan Huang, et al. "Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation?" In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 15184–15201.

[21] Michael Baumgartner, Paul F Jäger, Fabian Isensee, and Klaus H Maier-Hein. "nnDetection: a self-configuring method for medical object detection." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2021, pp. 530–539.

[22] Tatiana Bejarano et al. "Head-and-neck squamous cell carcinoma patients with CT taken during pre-treatment, mid-treatment, and post-treatment Dataset." In: *The Cancer Imaging Archive* 10 (2018), K9.

[23] Tatiana Bejarano et al. "Longitudinal fan-beam computed tomography dataset for head-and-neck squamous cell carcinoma patients." In: *Medical physics* 46.5 (2019), pp. 2526–2537.

[24] Alan Joseph Bekker and Jacob Goldberger. "Training deep neural-networks based on unreliable labels." In: *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, pp. 2682–2686.

[25] Aïcha BenTaieb and Ghassan Hamarneh. "Topology aware fully convolutional networks for histology gland segmentation." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 460–468.

[26] Cosmin I Bercea, Benedikt Wiestler, Daniel Rueckert, and Julia A Schnabel. "Diffusion models with implicit guidance for medical anomaly detection." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 211–220.

[27] Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved?" In: *IEEE transactions on medical imaging* 37.11 (2018), pp. 2514–2525.

[28] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. "The liver tumor segmentation benchmark (lits)." In: *Medical image analysis* 84 (2023), p. 102680.

[29] Risab Biswas. "Polyp-sam++: Can a text guided sam perform better for polyp segmentation?" In: *arXiv preprint arXiv:2308.06623* (2023).

[30] Quinn de Bourbon, Shadab Ahamed, Arman Rahmim, Paul Blanc-Durand, and Ran Klein. *Characterizing the limits of lesion detection by AI using synthetic lesions*. 2024.

[31] Yuri Y Boykov and M-P Jolly. "Interactive graph cuts for optimal boundary & region segmentation of objects in ND images." In: *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*. Vol. 1. IEEE. 2001, pp. 105–112.

[32] Gustav Bredell, Christine Tanner, and Ender Konukoglu. "Iterative interaction training for segmentation editing networks." In: *International workshop on machine learning in medical imaging*. Springer. 2018, pp. 363–370.

[33] Pete Bridge, Andrew Fielding, Pamela Rowntree, and Andrew Pullar. *Intraobserver variability: should we worry?* 2016.

[34] Christopher Brückner, Chang Liu, Leonhard Rist, and Andreas Maier. "Influence of imperfect annotations on deep learning segmentation models." In: *BVM Workshop*. Springer. 2024, pp. 226–231.

[35] Lishan Cai, Mohamed A Abdelatty, Luyi Han, Doenja MJ Lambregts, Joost van Griethuysen, Eduardo Pooch, Regina GH Beets-Tan, Sean Benson, Joren Brunekreef, and Jonas Teuwen. "Improving Rectal Tumor Segmentation with Anomaly Fusion Derived from Anatomical Inpainting: A Multicenter Study." In: *medRxiv* (2024), pp. 2024–10.

[36] Zhaowei Cai and Nuno Vasconcelos. "Cascade r-cnn: Delving into high quality object detection." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 6154–6162.

[37] Aaron Carass, Snehashis Roy, Adrian Gherman, Jacob C Reinhold, Andrew Jesson, Tal Arbel, Oskar Maier, Heinz Handels, Mohsen Ghafoorian, Bram Platel, et al. "Evaluating white matter lesion segmentations with refined Sørensen-Dice analysis." In: *Scientific reports* 10.1 (2020), p. 8242.

[38] M Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, et al. "Monai: An open-source framework for deep learning in healthcare." In: *arXiv preprint arXiv:2211.02701* (2022).

[39] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. "End-to-end object detection with transformers." In: *European conference on computer vision*. Springer. 2020, pp. 213–229.

[40] Shikha Chaganti, Philippe Grenier, Abishek Balachandran, Guillaume Chabin, Stuart Cohen, Thomas Flohr, Bogdan Georgescu, Sasa Grbic, Siqi Liu, François Mellot, et al. "Automated quantification of CT patterns associated with COVID-19 from chest CT." In: *Radiology: Artificial Intelligence* 2.4 (2020), e200048.

[41] Jayasree Chakraborty, Abhishek Midya, Lior Gazit, Marc Attiyeh, Liana Langdon-Embry, Peter J Allen, Richard KG Do, and Amber L Simpson. "CT radiomics to predict high-risk intraductal papillary mucinous neoplasms of the pancreas." In: *Medical physics* 45.11 (2018), pp. 5019–5029.

[42] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. "DCAN: deep contour-aware networks for accurate gland segmentation." In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 2016, pp. 2487–2496.

[43] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. "Encoder-decoder with atrous separable convolution for semantic image segmentation." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 801–818.

[44] Mingyang Chen, Yuting Wang, Qiankun Wang, Jingyi Shi, Huike Wang, Zichen Ye, Peng Xue, and Youlin Qiao. "Impact of human and artificial intelligence collaboration on workload reduction in medical image interpretation." In: *NPJ Digital Medicine* 7.1 (2024), p. 349.

[45] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. "Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more." In: *arXiv preprint arXiv:2304.09148* 2.5 (2023), p. 7.

[46] X. Chen, Z. Zhao, Y. Zhang, M. Duan, D. Qi, and H. Zhao. "FocalClick: Towards Practical Interactive Image Segmentation." In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, June 2022, pp. 1290–1299.

[47] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C Berg, and Alexander Kirillov. "Boundary IoU: Improving object-centric image segmentation evaluation." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 15334–15342.

[48] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. "Masked-attention mask transformer for universal image segmentation." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1290–1299.

[49] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. "Masked-attention Mask Transformer for Universal Image Segmentation." In: 2022.

[50] Dongjie Cheng, Ziyuan Qin, Zekun Jiang, Shaoting Zhang, Qicheng Lao, and Kang Li. "Sam on medical images: A comprehensive study on three prompt modes." In: *arXiv preprint arXiv:2305.00035* (2023).

[51] Sungduk Cho, Hyungjoon Jang, Jing Wei Tan, and Won-Ki Jeong. "Deepscribble: interactive pathology image segmentation using deep neural networks with scribbles." In: *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*. IEEE. 2021, pp. 761–765.

[52] Eric W. Christensen, Jay R. Parikh, Alexandra R. Drake, Eric M. Rubin, and Elizabeth Y. Rula. "Projected US Radiologist Supply, 2025 to 2055." In: *Journal of the American College of Radiology* 22.2 (Feb. 2025). Copyright © 2024 American College of Radiology. Published by Elsevier Inc. All rights reserved., pp. 161–169. ISSN: 1558-349X. DOI: 10.1016/j.jacr.2024.10.019. URL: https://doi.org/10.1016/j.jacr.2024.10.019.

[53] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 424–432.

[54] Kenneth Clark et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository." In: *Journal of digital imaging* 26 (2013), pp. 1045–1057.

[55] James R Clough, Nicholas Byrne, Ilkay Oksuz, Veronika A Zimmer, Julia A Schnabel, and Andrew P King. "A topological loss function for deep-learning based image segmentation using persistent homology." In: *IEEE transactions on pattern analysis and machine intelligence* 44.12 (2020), pp. 8766–8778.

[56] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)." In: *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 168–172.

[57] Adrian V Dalca, John Guttag, and Mert R Sabuncu. "Anatomical priors in convolutional networks for unsupervised biomedical segmentation." In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 9290–9299.

[58] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[59] A. Dervieux and F. Thomasset. "A finite element method for the simulation of a Rayleigh-Taylor instability." In: *Approximation Methods for Navier-Stokes Problems*. Ed. by Reimund Rautmann. Berlin, Heidelberg: Springer Berlin Heidelberg, 1980, pp. 145–158. ISBN: 978-3-540-38550-9.

[60] Yair Dgani, Hayit Greenspan, and Jacob Goldberger. "Training a neural network based on unreliable human annotation of medical images." In: *2018 IEEE 15th International symposium on biomedical imaging (ISBI 2018)*. IEEE. 2018, pp. 39–42.

[61] Lee R Dice. "Measures of the amount of ecologic association between species." In: *Ecology* 26.3 (1945), pp. 297–302.

[62] Huy M Do, Lillian G Spear, Moozhan Nikpanah, S Mojdeh Mirmomen, Laura B Machado, Alexandra P Toscano, Baris Turkbey, Mohammad Hadi Bagheri, James L Gulley, and Les R Folio. "Augmented radiologist workflow improves report value and saves time: a potential model for implementation of artificial intelligence." In: *Academic radiology* 27.1 (2020), pp. 96–105.

[63] Kaiqi Dong, Peijun Hu, Yu Tian, Yan Zhu, Xiang Li, Tianshu Zhou, Xueli Bai, Tingbo Liang, and Jingsong Li. "Position-aware representation learning with anatomical priors for enhanced pancreas tumor segmentation." In: *Neurocomputing* 616 (2025), p. 128881.

[64] Kaiqi Dong, Yan Zhu, Yu Tian, Peijun Hu, Chengkai Wu, Xiang Li, Tianshu Zhou, Xueli Bai, Tingbo Liang, and Jingsong Li. "A Knowledge-Driven Evidence Fusion Network for pancreatic tumor segmentation in CT images." In: *Biomedical Signal Processing and Control* 111 (2026), p. 108281.

[65] Olaf Dössel. *Bildgebende Verfahren in der Medizin*. Berlin, Heidelberg: Springer, 2016.

[66] Audrey Duran, Gaspard Dussert, Olivier Rouvière, Tristan Jaouen, Pierre-Marc Jodoin, and Carole Lartizien. "ProstAttention-Net: A deep attention model for prostate cancer segmentation by aggressiveness in MRI scans." In: *Medical Image Analysis* 77 (2022), p. 102347.

[67] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. "The pascal visual object classes (voc) challenge." In: *International journal of computer vision* 88.2 (2010), pp. 303–338.

[68] William A Falcon. *Pytorch lightning*. 2019.

[69] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. "Inf-net: Automatic covid-19 lung infection segmentation from ct images." In: *IEEE transactions on medical imaging* 39.8 (2020), pp. 2626–2637.

[70] Chaowei Fang, Qian Wang, Lechao Cheng, Zhifan Gao, Chengwei Pan, Zhen Cao, Zhaohui Zheng, and Dingwen Zhang. "Reliable mutual distillation for medical image segmentation under imperfect annotations." In: *IEEE Transactions on Medical Imaging* 42.6 (2023), pp. 1720–1734.

[71] Adrien Foucart, Olivier Debeir, and Christine Decaestecker. "Panoptic quality should be avoided as a metric for assessing cell nuclei segmentation and classification in digital pathology." In: *Scientific reports* 13.1 (2023), p. 8614.

[72] Daniel Freedman and Tao Zhang. "Interactive graph cut based segmentation with shape priors." In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1. IEEE. 2005, pp. 755–762.

[73] Xiaohang Fu, Lei Bi, Ashnil Kumar, Michael Fulham, and Jinman Kim. "Multimodal spatial attention module for targeting multimodal PET-CT lung tumor segmentation." In: *IEEE Journal of Biomedical and Health Informatics* 25.9 (2021), pp. 3507–3516.

[74] Gaëtan Galisot, Jean-Yves Ramel, Thierry Brouard, Elodie Chaillou, and Barthélémy Serres. "Visual and structural feature combination in an interactive machine learning system for medical image segmentation." In: *Machine Learning with Applications* 8 (2022), p. 100294.

[75] Mostafa Gamal, Mennatullah Siam, and Moemen Abdel-Razek. "Shuffleseg: Real-time semantic segmentation network." In: *arXiv preprint arXiv:1803.03816* (2018).

[76] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. "Clip-adapter: Better vision-language models with feature adapters." In: *International Journal of Computer Vision* 132.2 (2024), pp. 581–595.

[77] Shengbo Gao, Ziji Zhang, Jiechao Ma, Zihao Li, and Shu Zhang. "Correlation-aware mutual learning for semi-supervised medical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 98–108.

[78] Yunhe Gao, Difei Gu, Mu Zhou, and Dimitris Metaxas. "Aligning human knowledge with visual concepts towards explainable medical image classification." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 46–56.

[79] Yunhe Gao, Rui Huang, Yiwei Yang, Jie Zhang, Kainan Shao, Changjuan Tao, Yuanyuan Chen, Dimitris N Metaxas, Hongsheng Li, and Ming Chen. "FocusNetv2: Imbalanced large and small organ segmentation with adversarial shape constraint for head and neck CT images." In: *Medical Image Analysis* 67 (2021), p. 101831.

[80] Sergios Gatidis, Marcel Früh, Matthias P Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, et al. "Results from the autoPET challenge on fully automated lesion segmentation in oncologic PET/CT imaging." In: *Nature Machine Intelligence* 6.11 (2024), pp. 1396–1405.

[81] Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. "A whole-body FDG-PET/CT dataset with manually annotated tumor lesions." In: *Scientific Data* 9.1 (2022), p. 601.

[82] J Raymond Geis, Adrian P Brady, Carol C Wu, Jack Spencer, Erik Ranschaert, Jacob L Jaremko, Steve G Langer, Andrea Borondy Kitts, Judy Birch, William F Shields, et al. "Ethics of artificial intelligence in radiology: summary of the joint European and North American multisociety statement." In: *Radiology* 293.2 (2019), pp. 436–440.

[83] Amirata Ghorbani, David Ouyang, Abubakar Abid, Bryan He, Jonathan H Chen, Robert A Harrington, David H Liang, Euan A Ashley, and James Y Zou. "Deep learning interpretation of echocardiograms." In: *NPJ digital medicine* 3.1 (2020), p. 10.

[84] Aritra Ghosh, Himanshu Kumar, and P Shanti Sastry. "Robust loss functions under label noise for deep neural networks." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 31. 1. 2017.

[85] Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. "Automatic multi-organ segmentation on abdominal CT with dense V-networks." In: *IEEE transactions on medical imaging* 37.8 (2018), pp. 1822–1834.

[86] Kristina Giske et al. "Local setup errors in image-guided radiotherapy for head and neck cancer patients immobilized with a custom-made device." In: *International Journal of Radiation Oncology* Biology* Physics* 80.2 (2011), pp. 582–589.

[87] Rotem Golan, Christian Jacob, and Jörg Denzinger. "Lung nodule detection in CT images using deep convolutional neural networks." In: *2016 international joint conference on neural networks (IJCNN)*. IEEE. 2016, pp. 243–250.

[88] Shizhan Gong, Yuan Zhong, Wenao Ma, Jinpeng Li, Zhao Wang, Jingyang Zhang, Pheng-Ann Heng, and Qi Dou. "3dsam-adapter: Holistic adaptation of sam from 2d to 3d for promptable tumor segmentation." In: *Medical Image Analysis* 98 (2024), p. 103324.

[89] Alvaro Gonzalez-Jimenez, Simone Lionetti, Philippe Gottfrois, Fabian Gröger, Marc Pouly, and Alexander A Navarini. "Robust t-loss for medical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 714–724.

[90] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. "The llama 3 herd of models." In: *arXiv preprint arXiv:2407.21783* (2024).

[91] Vincent Grégoire, Kian Ang, Wilfried Budach, Cai Grau, Marc Hamoir, Johannes A Langendijk, Anne Lee, Quynh-Thu Le, Philippe Maingon, Chris Nutting, et al. "Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines." In: *Radiotherapy and Oncology* 110.1 (2014), pp. 172–181.

[92] Vincent Grégoire, Mererid Evans, Quynh-Thu Le, Jean Bourhis, Volker Budach, Amy Chen, Abraham Eisbruch, Mei Feng, Jordi Giralt, Tejpal Gupta, et al. "Delineation of the primary tumour clinical target volumes (ctv-p) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: Airo, caca, dahanca, eortc, georcc, gortec, hknpcsg, hncig, iag-kht, lprhht, ncic ctg, ncri, nrg oncology, phns, sbrt, somera, sro, sshno, trog consensus guidelines." In: *Radiotherapy and Oncology* 126.1 (2018), pp. 3–24.

[93] Mahdi Hajiaghayi, Elliott M Groves, Hamid Jafarkhani, and Arash Kheradvar. "A 3-D active contour method for automated segmentation of the left ventricle from magnetic resonance images." In: *IEEE Transactions on Biomedical Engineering* 64.1 (2016), pp. 134–144.

[94] Alexander Hammers, Richard Allom, Matthias J Koepp, Samantha L Free, Ralph Myers, Louis Lemieux, Tejal N Mitchell, David J Brooks, and John S Duncan. "Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe." In: *Human brain mapping* 19.4 (2003), pp. 224–247.

[95] Kai Han, Lu Liu, Yuqing Song, Yi Liu, Chengjian Qiu, Yangyang Tang, Qiaoying Teng, and Zhe Liu. "An effective semi-supervised approach for liver CT image segmentation." In: *IEEE Journal of Biomedical and Health Informatics* 26.8 (2022), pp. 3999–4007.

[96] Kai Han, Victor S Sheng, Yuqing Song, Yi Liu, Chengjian Qiu, Siqi Ma, and Zhe Liu. "Deep semi-supervised learning for medical image segmentation: A review." In: *Expert Systems with Applications* 245 (2024), p. 123052.

[97] Fahmida Haque, Alex Chen, Nathan Lay, Jorge Carrasquillo, Esther Mena, Julia Segal, Philip Eclarinal, Peter Choyke, Rosandra Kaplan, Frank Lin, et al. *Development and validation of pan-cancer lesion segmentation AI-model for whole-body FDG PET/CT in diverse clinical cohorts*. 2024.

[98] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images." In: (Jan. 2022). arXiv: 2201.01266. URL: http://arxiv.org/abs/2201.01266.

[99] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. "Unetr: Transformers for 3d medical image segmentation." In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2022, pp. 574–584.

[100] Felix Hausdorff. *Grundzuge der mengenlehre*. Vol. 61. American Mathematical Soc., 1978.

[101] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. "Mask r-cnn." In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2961–2969.

[102] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[103] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. "VISTA3D: A unified segmentation foundation model for 3D medical imaging." In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 20863–20873.

[104] Tobias Heimann, Bram Van Ginneken, Martin A Styner, Yulia Arzhaeva, Volker Aurich, Christian Bauer, Andreas Beck, Christoph Becker, Reinhard Beichel, György Bekes, et al. "Comparison and evaluation of methods for liver segmentation from CT datasets." In: *IEEE transactions on medical imaging* 28.8 (2009), pp. 1251–1265.

[105] Lena Heinemann*, Alexander Jaus*, Zdravko Marinov, Moon Kim, Maria Francesca Spadea, Jens Kleesiek, and Rainer Stiefelhagen. "LIMIS: Towards Language-Based Interactive Medical Image Segmentation." In: *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2025, pp. 1–5.

[106] Nicholas Heller, Fabian Isensee, Klaus H Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, et al. "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 Challenge." In: *Medical Image Analysis* (2020), p. 101821.

[107] Nicholas Heller et al. *The KiTS21 Challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT*. 2023. arXiv: 2307.01984 [cs.CV].

[108] Ward Hendrix, Nils Hendrix, Ernst T Scholten, Mariëlle Mourits, Joline Trap-de Jong, Steven Schalekamp, Mike Korst, Maarten Van Leuken, Bram Van Ginneken, Mathias Prokop, et al. "Deep learning for the detection of benign and malignant pulmonary nodules in non-screening chest CT scans." In: *Communications medicine* 3.1 (2023), p. 156.

[109] James M Hillis, Jacob J Visser, Edward R Scheffer Cliff, Kelly van der Geest–Aspers, Bernardo C Bizzo, Keith J Dreyer, Jeremias Adams-Prassl, and Katherine P Andriole. "The lucent yet opaque challenge of regulating artificial intelligence in radiology." In: *NPJ Digital Medicine* 7.1 (2024), p. 69.

[110] Geoffrey Hinton. *Machine Learning and the Market for Intelligence – Panel Discussion*. Quote at 00:42–00:51. Creative Destruction Lab. Nov. 2016. URL: https://youtu.be/2HMPRXstSvQ (visited on 06/09/2025).

[111] Lukas Hirsch, Yu Huang, Shaojun Luo, Carolina Rossi Saccarelli, Roberto Lo Gullo, Isaac Daimiel Naranjo, Almir GV Bitencourt, Natsuko Onishi, Eun Sook Ko, Doris Leithner, et al. "Radiologist-level performance by using deep learning for segmentation of breast cancers on MRI scans." In: *Radiology: Artificial Intelligence* 4.1 (2021), e200231.

[112] Andrew Hoopes, Victor Ion Butoi, John V Guttag, and Adrian V Dalca. "Voxelprompt: A vision-language agent for grounded medical image analysis." In: *arXiv preprint arXiv:2410.08397* (2024).

[113] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. "Lora: Low-rank adaptation of large language models." In: *ICLR* 1.2 (2022), p. 3.

[114] Jie Hu, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7132–7141.

[115] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. "Label-free liver tumor segmentation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 7422–7432.

[116] Yuzhou Hu, Yi Guo, Yuanyuan Wang, Jinhua Yu, Jiawei Li, Shichong Zhou, and Cai Chang. "Automatic tumor segmentation in breast ultrasound images using a dilated fully convolutional network combined with an active contour model." In: *Medical physics* 46.1 (2019), pp. 215–228.

[117] Huimin Huang, Han Zheng, Lanfen Lin, Ming Cai, Hongjie Hu, Qiaowei Zhang, Qingqing Chen, Yutaro Iwamoto, Xianhua Han, Yen-Wei Chen, et al. "Medical image segmentation with deep atlas prior." In: *IEEE Transactions on Medical Imaging* 40.12 (2021), pp. 3519–3530.

[118] Ziyan Huang, Haoyu Wang, Zhongying Deng, Jin Ye, Yanzhou Su, Hui Sun, Junjun He, Yun Gu, Lixu Gu, Shaoting Zhang, et al. "Stu-net: Scalable and transferable medical image segmentation models empowered by large-scale supervised pre-training." In: *arXiv preprint arXiv:2304.06716* (2023).

[119] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 590–597.

[120] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation." In: *Nature methods* 18.2 (2021), pp. 203–211.

[121] Fabian Isensee, Paul F Jäger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. "Automated design of deep learning methods for biomedical image segmentation." In: *arXiv preprint arXiv:1904.08128* (2019).

[122] Fabian Isensee, Maximilian Rokuss, Lars Krämer, Stefan Dinkelacker, Ashis Ravindran, Florian Stritzke, Benjamin Hamm, Tassilo Wald, Moritz Langenberg, Constantin Ulrich, et al. "nninteractive: Redefining 3d promptable segmentation." In: *arXiv preprint arXiv:2503.08373* (2025).

[123] Mobarakol Islam and Ben Glocker. "Spatially varying label smoothing: Capturing uncertainty from expert annotations." In: *international conference on information processing in medical imaging*. Springer. 2021, pp. 677–688.

[124] Paul Jaccard. "The distribution of the flora in the alpine zone. 1." In: *New phytologist* 11.2 (1912), pp. 37–50.

[125] Paul F Jaeger, Simon AA Kohl, Sebastian Bickelhaupt, Fabian Isensee, Tristan Anselm Kuder, Heinz-Peter Schlemmer, and Klaus H Maier-Hein. "Retina U-Net: Embarrassingly simple exploitation of segmentation supervision for medical object detection." In: *Machine learning for health workshop*. PMLR. 2020, pp. 171–183.

[126] Stefan Jaeger, Sinan Candemir, Sameer Antani, Yifan Wang, Pu-Xuan Lu, and George Thoma. “Two public chest X-ray datasets for computer-aided screening of pulmonary diseases.” In: *Quantitative imaging in medicine and surgery* 4.6 (2014), pp. 475–477.

[127] Alexander Jaus, Zdravko Marinov, Constantin Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. “Good Enough: Is it Worth Improving your Label Quality?” In: *arXiv preprint arXiv:2505.20928* (2025).

[128] Alexander Jaus, Constantin Seibold, Simon Reiß, Lukas Heine, Anton Schily, Moon Kim, Fin Hendrik Bahnsen, Ken Herrmann, Rainer Stiefelhagen, and Jens Kleesiek. “Anatomy-guided Pathology Segmentation.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer Nature Switzerland Cham. 2024, pp. 3–13.

[129] Alexander Jaus, Constantin Marc Seibold, Simon Reiß, Zdravko Marinov, Keyi Li, Zeling Ye, Stefan Krieg, Jens Kleesiek, and Rainer Stiefelhagen. “Every Component Counts: Rethinking the Measure of Success for Medical Semantic Segmentation in Multi-Instance Segmentation Tasks.” In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 4. 2025, pp. 3904–3912.

[130] Alexander Jaus*, Constantin Seibold*, Kelsey Hermann, Negar Shahamiri, Alexandra Walter, Kristina Giske, Johannes Haubold, Jens Kleesiek, and Rainer Stiefelhagen. “Towards Unifying Anatomy Segmentation: Automated Generation of a Full-Body CT Dataset.” In: *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE. 2024, pp. 41–47.

[131] Alexander Jaus*, Constantin Seibold*, Kelsey Hermann, Alexandra Walter, Kristina Giske, Johannes Haubold, Jens Kleesiek, and Rainer Stiefelhagen. “Towards unifying anatomy segmentation: automated generation of a full-body ct dataset via knowledge aggregation and anatomical guidelines.” In: *arXiv preprint arXiv:2307.13375* (2023).

[132] Young Seok Jeon, Hongfei Yang, Huazhu Fu, and Mengling Feng. “Teaching AI the Anatomy Behind the Scan: Addressing Anatomical Flaws in Medical Image Segmentation with Learnable Prior.” In: *arXiv preprint arXiv:2403.18878* (2024).

[133] Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. “Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation.” In: *Advances in neural information processing systems* 35 (2022), pp. 36722–36732.

[134] Juntao Jiang, Xiyu Chen, Guanzhong Tian, and Yong Liu. “ViG-UNet: vision graph neural networks for medical image segmentation.” In: *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2023, pp. 1–5.

[135] Yankai Jiang, Peng Zhang, Donglin Yang, Yuan Tian, Hai Lin, and Xiaosong Wang. "Advancing Generalizable Tumor Segmentation with Anomaly-Aware Open-Vocabulary Attention Maps and Frozen Foundation Diffusion Models." In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 25971–25981.

[136] Oscar Jimenez-del-Toro, Henning Müller, Markus Krenn, Katharina Gruenberg, Abdel Aziz Taha, Marianne Winterstein, Ivan Eggel, Antonio Foncubierta-Rodríguez, Orcun Goksel, András Jakab, et al. "Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks." In: *IEEE transactions on medical imaging* 35.11 (2016), pp. 2459–2475.

[137] Liang Jin, Shixuan Gu, Donglai Wei, Jason Ken Adhinarta, Kaiming Kuang, Yongjie Jessica Zhang, Hanspeter Pfister, Bingbing Ni, Jiancheng Yang, and Ming Li. "RibSeg v2: A Large-scale Benchmark for Rib Labeling and Anatomical Centerline Extraction." In: *IEEE Transactions on Medical Imaging (TMI)* (2023).

[138] Liang Jin, Jiancheng Yang, Kaiming Kuang, Bingbing Ni, Yiyi Gao, Yingli Sun, Pan Gao, Weiling Ma, Mingyu Tan, Hui Kang, et al. "Deep-learning-assisted detection and segmentation of rib fractures from CT scans: Development and validation of FracNet." In: *EBioMedicine* 62 (2020), p. 103106.

[139] Petr Jordan, Philip M Adamson, Vrunda Bhattbhatt, Surabhi Beriwal, Sangyu Shen, Oskar Radermecker, Supratik Bose, Linda S Strain, Michael Offe, David Fraley, et al. "Pediatric chest-abdomen-pelvis and abdomen-pelvis CT images with expert organ contours." In: *Medical physics* 49.5 (2022), pp. 3523–3528.

[140] Lie Ju, Xin Wang, Lin Wang, Dwarikanath Mahapatra, Xin Zhao, Quan Zhou, Tongliang Liu, and Zongyuan Ge. "Improving medical images classification with label noise using dual-uncertainty estimation." In: *IEEE transactions on medical imaging* 41.6 (2022), pp. 1533–1546.

[141] Hamza Kalisch, Fabian Hörst, Ken Herrmann, Jens Kleesiek, and Constantin Seibold. "Autopet III challenge: Incorporating anatomical knowledge into nnUNet for lesion segmentation in PET/CT." In: *arXiv preprint arXiv:2409.12155* (2024).

[142] Ming Kang, Chee-Ming Ting, Fung Fung Ting, and Raphaël C-W Phan. "RCS-YOLO: A fast and high-accuracy object detector for brain tumor detection." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2023, pp. 600–610.

[143] Ming Kang, Chee-Ming Ting, Fung Fung Ting, and Raphaël C-W Phan. "Bgf-yolo: Enhanced yolov8 with multiscale attentional feature fusion for brain tumor detection." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 35–45.

[144] Davood Karimi, Haoran Dou, Simon K Warfield, and Ali Gholipour. "Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis." In: *Medical image analysis* 65 (2020), p. 101759.

[145] Davood Karimi, Caitlin K Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, and Ali Gholipour. "Learning to segment fetal brain tissue from noisy annotations." In: *Medical image analysis* 85 (2023), p. 102731.

[146] Michael Kass, Andrew Witkin, and Demetri Terzopoulos. "Snakes: Active contour models." In: *International journal of computer vision* 1.4 (1988), pp. 321–331.

[147] Prabhpreet Kaur, Gurvinder Singh, and Parminder Kaur. "A review of denoising medical images using machine learning approaches." In: *Current medical imaging* 14.5 (2018), pp. 675–685.

[148] A Emre Kavur, N Sinem Gezer, Mustafa Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. "CHAOS challenge-combined (CT-MR) healthy abdominal organ segmentation." In: *Medical image analysis* 69 (2021), p. 101950.

[149] Karsten Keller, Christoph Sinning, Andreas Schulz, Claus Jünger, Volker H Schmitt, Omar Hahad, Tanja Zeller, Manfred Beutel, Norbert Pfeiffer, Konstantin Strauch, et al. "Right atrium size in the general population." In: *Scientific reports* 11.1 (2021), p. 22523.

[150] Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, and Ismail Ben Ayed. "Constrained-CNN losses for weakly supervised segmentation." In: *Medical image analysis* 54 (2019), pp. 88–99.

[151] Ron Kikinis, Steve D Pieper, and Kirby G Vosburgh. "3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support." In: *Intraoperative imaging and image-guided therapy*. Springer, 2013, pp. 277–289.

[152] Yannick Kirchhoff, Maximilian R Rokuss, Saikat Roy, Balint Kovacs, Constantin Ulrich, Tassilo Wald, Maximilian Zenk, Philipp Vollmuth, Jens Kleesiek, Fabian Isensee, et al. "Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures." In: *European Conference on Computer Vision*. Springer. 2024, pp. 218–234.

[153] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. "Panoptic segmentation." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 9404–9413.

[154] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. "Segment anything." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 4015–4026.

[155] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. "Pointrend: Image segmentation as rendering." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9799–9808.

[156] Sven Koitka et al. "Fully automated body composition analysis in routine CT imaging using 3D semantic segmentation convolutional neural networks." In: *European radiology* 31 (2021), pp. 1795–1804.

[157] Sven Koitka, Giulia Baldini, Lennard Kroll, Natalie van Landeghem, Olivia B Pollok, Johannes Haubold, Obioma Pelka, Moon Kim, Jens Kleesiek, Felix Nensa, et al. "SAROS: A dataset for whole-body region and organ segmentation in CT imaging." In: *Scientific Data* 11.1 (2024), p. 483.

[158] Taha Koleilat, Hojat Asgariandehkordi, Hassan Rivaz, and Yiming Xiao. "Medclip-samv2: Towards universal text-driven medical image segmentation." In: *arXiv preprint arXiv:2409.19483* (2024).

[159] Navid Alemi Koohbanani, Mostafa Jahanifar, Neda Zamani Tajadin, and Nasir Rajpoot. "NuClick: a deep learning framework for interactive segmentation of microscopic images." In: *Medical Image Analysis* 65 (2020), p. 101771.

[160] Hans-Peter Kriegel, Peer Kröger, Erich Schubert, and Arthur Zimek. "LoOP: local outlier probabilities." In: *Proceedings of the 18th ACM conference on Information and knowledge management*. 2009, pp. 1649–1652.

[161] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. "Visual genome: Connecting language and vision using crowdsourced dense image annotations." In: *International journal of computer vision* 123 (2017), pp. 32–73.

[162] Anitha Priya Krishnan, Zhuang Song, David Clayton, Laura Gaetano, Xiaoming Jia, Alex de Crespigny, Thomas Bengtsson, and Richard AD Carano. "Joint MRI T1 unenhancing and contrast-enhancing multiple sclerosis lesion segmentation with deep learning in OPERA trials." In: *Radiology* 302.3 (2022), pp. 662–673.

[163] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks." In: *Advances in neural information processing systems* 25 (2012).

[164] Abhishek Kumar and Hal Daumé. "A co-training approach for multi-view spectral clustering." In: *Proceedings of the 28th international conference on machine learning (ICML-11)*. 2011, pp. 393–400.

[165] Vishal Kumar, Vishnu Baburaj, Sandeep Patel, Siddhartha Sharma, and Raju Vaishya. "Does the use of intraoperative CT scan improve outcomes in Orthopaedic surgery? A systematic review and meta-analysis of 871 cases." In: *Journal of Clinical Orthopaedics and Trauma* 18 (2021), pp. 216–223.

[166] Thomas Küstner, Tobias Hepp, Marc Fischer, Martin Schwartz, Andreas Fritsche, Hans-Ulrich Häring, Konstantin Nikolaou, Fabian Bamberg, Bin Yang, Fritz Schick, et al. "Fully automated and standardized segmentation of adipose tissue compartments via deep learning in 3D whole-body MRI of epidemiologic cohort studies." In: *Radiology: Artificial Intelligence* 2.6 (2020), e200010.

[167] Zoé Lambert, Caroline Petitjean, Bernard Dubray, and Su Kuan. "Segthor: Segmentation of thoracic organs at risk in ct images." In: *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*. IEEE. 2020, pp. 1–6.

[168] Bennett Landman, Zhoubing Xu, J Igelsias, Martin Styner, T Langerak, and Arno Klein. "Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge." In: *Proc. MICCAI Multi-Atlas Labeling Beyond Cranial Vault—Workshop Challenge*. Vol. 5. 2015, p. 12.

[169] Kristina Lång, Viktoria Josefsson, Anna-Maria Larsson, Stefan Larsson, Charlotte Högberg, Hanna Sartor, Solveig Hofvind, Ingvar Andersson, and Aldana Rosso. "Artificial intelligence-supported screen reading versus standard double reading in the Mammography Screening with Artificial Intelligence trial (MASAI): a clinical safety analysis of a randomised, controlled, non-inferiority, single-blinded, screening accuracy study." In: *The Lancet Oncology* 24.8 (2023), pp. 936–944.

[170] Jan Larsen, L Nonboe, Mads Hintz-Madsen, and Lars Kai Hansen. "Design of robust neural network classifiers." In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*. Vol. 2. IEEE. 1998, pp. 1205–1208.

[171] Christopher P Lee, Zhoubing Xu, Ryan P Burke, Rebeccah B Baucom, Benjamin K Poulose, Richard G Abramson, and Bennett A Landman. "Evaluation of five image registration tools for abdominal CT: pitfalls and opportunities with soft anatomy." In: *Proceedings of Spie–the International Society for Optical Engineering*. Vol. 9413. 2015, 94131N.

[172] Hyunkwang Lee, Sehyo Yune, Mohammad Mansouri, Myeongchan Kim, Shahein H Tajmir, Claude E Guerrier, Sarah A Ebert, Stuart R Pomerantz, Javier M Romero, Shahmir Kamalian, et al. "An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets." In: *Nature biomedical engineering* 3.3 (2019), pp. 173–182.

[173] Wenhui Lei, Huan Wang, Ran Gu, Shichuan Zhang, Shaoting Zhang, and Guotai Wang. "DeepIGeoS-V2: Deep interactive segmentation of multiple organs from head and neck images with lightweight CNNs." In: *International Workshop on Large-scale Annotation of Biomedical data and Expert Label Synthesis*. Springer. 2019, pp. 61–69.

[174] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. "Llava-med: Training a large language-and-vision assistant for biomedicine in one day." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 28541–28564.

[175] Feng Li, Hao Zhang, Huaizhe Xu, Shilong Liu, Lei Zhang, Lionel M Ni, and Heung-Yeung Shum. "Mask dino: Towards a unified transformer-based framework for object detection and segmentation." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 3041–3050.

[176] Lin Li, Lixin Qin, Zeguo Xu, Youbing Yin, Xin Wang, Bin Kong, Junjie Bai, Yi Lu, Zhenghan Fang, Qi Song, et al. "Using artificial intelligence to detect COVID-19 and community-acquired pneumonia based on pulmonary CT: evaluation of the diagnostic accuracy." In: *Radiology* 296.2 (2020), E65–E71.

[177] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, et al. "AbdomenAtlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking." In: *Medical Image Analysis* (2024), p. 103285. URL: https://github.com/MrGiovanni/AbdomenAtlas.

[178] Wenxuan Li, Alan Yuille, and Zongwei Zhou. "How Well Do Supervised 3D Models Transfer to Medical Imaging Tasks?" In: *The Twelfth International Conference on Learning Representations*. 2024. URL: https://openreview.net/forum?id=AhizIPytk4.

[179] Wenxuan Li, Alan Yuille, and Zongwei Zhou. "How Well Do Supervised Models Transfer to 3D Image Segmentation?" In: *The Twelfth International Conference on Learning Representations*. 2024.

[180] Keyi Li*, Alexander Jaus*, Jens Kleesiek, and Rainer Stiefelhagen. "GRASPing Anatomy to Improve Pathology Segmentation." In: *International Workshop on Machine Learning in Medical Imaging*. Springer. 2025.

[181] Xuan Liao, Wenhao Li, Qisen Xu, Xiangfeng Wang, Bo Jin, Xiaoyun Zhang, Yanfeng Wang, and Ya Zhang. "Iteratively-refined interactive 3D medical image segmentation with multi-agent reinforcement learning." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 9394–9402.

[182] Zehui Liao, Yutong Xie, Shishuai Hu, and Yong Xia. "Learning from ambiguous labels for lung nodule malignancy prediction." In: *IEEE Transactions on medical imaging* 41.7 (2022), pp. 1874–1884.

[183] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. "Microsoft coco: Common objects in context." In: *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer. 2014, pp. 740–755.

[184] Zheng Lin, Zhao Zhang, Ling-Hao Han, and Shao-Ping Lu. "Multi-mode interactive image segmentation." In: *Proceedings of the 30th ACM International Conference on Multimedia*. 2022, pp. 905–914.

[185] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. "Clip-driven universal model for organ segmentation and tumor detection." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 21152–21164.

[186] Jingyu Liu, Jie Lian, and Yizhou Yu. *ChestX-Det10: Chest X-ray Dataset on Detection of Thoracic Abnormalities*. 2020. arXiv: 2006.10550 [eess.IV].

[187] Pengbo Liu, Hu Han, Yuanqi Du, Heqin Zhu, Yinhao Li, Feng Gu, Honghu Xiao, Jun Li, Chunpeng Zhao, Li Xiao, et al. "Deep learning to segment pelvic bones: large-scale CT datasets and baseline models." In: *International Journal of Computer Assisted Radiology and Surgery* 16 (2021), pp. 749–756.

[188] Qin Liu, Zhenlin Xu, Yining Jiao, and Marc Niethammer. "isegformer: interactive segmentation via transformers with application to 3d knee mr images." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2022, pp. 464–474.

[189] Sheng Liu et al. "Early-learning regularization prevents memorization of noisy labels." In: *Advances in neural information processing systems* 33 (2020), pp. 20331–20342.

[190] Sheng Liu, Kangning Liu, Weicheng Zhu, Yiqiu Shen, and Carlos Fernandez-Granda. "Adaptive early-learning correction for segmentation from noisy annotations." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 2606–2616.

[191] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. "Grounding dino: Marrying dino with grounded pre-training for open-set object detection." In: *European conference on computer vision*. Springer. 2024, pp. 38–55.

[192] Yanzhen Liu, Sutuke Yibulayimu, Yudi Sang, Gang Zhu, Chao Shi, Chendi Liang, Qiyong Cao, Chunpeng Zhao, Xinbao Wu, and Yu Wang. "Preoperative fracture reduction planning for image-guided pelvic trauma surgery: A comprehensive pipeline with learning." In: *Medical Image Analysis* 102 (2025), p. 103506.

ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2025.103506. URL: https://www.sciencedirect.com/science/article/pii/S1361841525000544.

[193] Yanzhen Liu, Sutuke Yibulayimu, Gang Zhu, Chao Shi, Chendi Liang, Chunpeng Zhao, Xinbao Wu, Yudi Sang, and Yu Wang. "Automatic pelvic fracture segmentation: a deep learning approach and benchmark dataset." In: *Frontiers in Medicine* 12 (2025), p. 1511487.

[194] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. "A convnet for the 2020s." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11976–11986.

[195] Pechin Lo, Bram Van Ginneken, Joseph M Reinhardt, Tarunashree Yavarna, Pim A De Jong, Benjamin Irving, Catalin Fetita, Margarete Ortner, Rômulo Pinho, Jan Sijbers, et al. "Extraction of airways from CT (EXACT'09)." In: *IEEE Transactions on Medical Imaging* 31.11 (2012), pp. 2093–2107.

[196] *LObe and Lung Analysis 2011 (LOLA11)*. https://lola11.grand-challenge.org/. Grand Challenge dataset for lung and lobe segmentation. 2011.

[197] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[198] Ilya Loshchilov and Frank Hutter. "SGDR: Stochastic Gradient Descent with Warm Restarts." In: *arXiv preprint arXiv:1608.03983* (2016).

[199] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization." In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[200] J John Lucido, Todd A DeWees, Todd R Leavitt, Aman Anand, Chris J Beltran, Mark D Brooke, Justine R Buroker, Robert L Foote, Olivia R Foss, Angela M Gleason, et al. "Validation of clinical acceptability of deep-learning-based automated segmentation of organs-at-risk for head-and-neck radiotherapy treatment planning." In: *Frontiers in Oncology* 13 (2023), p. 1137803.

[201] Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, et al. "Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge." In: *arXiv preprint arXiv:2304.03708* (2023).

[202] Xiangde Luo, Wenjun Liao, Jianghong Xiao, Jieneng Chen, Tao Song, Xiaofan Zhang, Kang Li, Dimitris N Metaxas, Guotai Wang, and Shaoting Zhang. "WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image." In: *Medical Image Analysis* 82 (2022), p. 102642.

[203] Fei Lyu, Mang Ye, Jonathan Frederik Carlsen, Kenny Erleben, Sune Darkner, and Pong C Yuen. "Pseudo-label guided image synthesis for semi-supervised covid-19 pneumonia infection segmentation." In: *IEEE Transactions on Medical Imaging* 42.3 (2022), pp. 797–809.

[204] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. "Segment anything in medical images." In: *Nature Communications* 15.1 (2024), p. 654.

[205] Jun Ma, Yao Zhang, Song Gu, Cheng Ge, Shihao Ma, Adamo Young, Cheng Zhu, Kangkang Meng, Xin Yang, Ziyan Huang, et al. "Unleashing the strengths of unlabeled data in pan-cancer abdominal organ quantification: the flare22 challenge." In: *arXiv preprint arXiv:2308.05862* (2023).

[206] Jun Ma, Yao Zhang, Song Gu, Cheng Zhu, Cheng Ge, Yichi Zhang, Xingle An, Congcong Wang, Qiyuan Wang, Xin Liu, et al. "Abdomenct-1k: Is abdominal organ segmentation a solved problem?" In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.10 (2021), pp. 6695–6714.

[207] Jun Ma et al. "Fast and Low-GPU-memory abdomen CT organ segmentation: The FLARE challenge." In: *Medical Image Analysis* 82 (2022), p. 102616.

[208] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. "Iteratively trained interactive segmentation." In: *arXiv preprint arXiv:1805.04398* (2018).

[209] Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, et al. "Metrics reloaded: recommendations for image analysis validation." In: *Nature methods* 21.2 (2024), pp. 195–212.

[210] Zdravko Marinov, Paul F Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen. "Deep interactive segmentation of medical images: A systematic review and taxonomy." In: *IEEE transactions on pattern analysis and machine intelligence* (2024).

[211] Zdravko Marinov, Moon Kim, Jens Kleesiek, and Rainer Stiefelhagen. "Rethinking Annotator Simulation: Realistic Evaluation of Whole-Body PET Lesion Interactive Segmentation Methods." In: *arXiv preprint arXiv:2404.01816* (2024).

[212] Zdravko Marinov, Simon Reiß, David Kersting, Jens Kleesiek, and Rainer Stiefelhagen. "Mirror u-net: Marrying multimodal fission with multi-task learning for semantic segmentation in medical imaging." In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 2283–2293.

[213] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. "Segment anything model for medical image analysis: an experimental study." In: *Medical Image Analysis* 89 (2023), p. 102918.

[214] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. "International evaluation of an AI system for breast cancer screening." In: *Nature* 577.7788 (2020), pp. 89–94.

[215] Peter McLaughlin, Lee Benson, and Eric Horlick. "The Role of Cardiac Catheterization in Adult Congenital Heart Disease." In: *Cardiology Clinics* 24.4 (2006), pp. 531–556.

[216] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)." In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[217] Duy MH Nguyen, Hoang Nguyen, Nghiem Diep, Tan Ngoc Pham, Tri Cao, Binh Nguyen, Paul Swoboda, Nhat Ho, Shadi Albarqouni, Pengtao Xie, et al. "Lvm-med: Learning large-scale self-supervised vision models for medical imaging via second-order graph matching." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 27922–27950.

[218] Juzheng Miao, Cheng Chen, Furui Liu, Hao Wei, and Pheng-Ann Heng. "Caussl: Causality-inspired semi-supervised learning for medical image segmentation." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 21426–21437.

[219] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation." In: *2016 fourth international conference on 3D vision (3DV)*. Ieee. 2016, pp. 565–571.

[220] Shaobo Min, Xuejin Chen, Zheng-Jun Zha, Feng Wu, and Yongdong Zhang. "A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01. 2019, pp. 4578–4585.

[221] Zahra Mirikharaji and Ghassan Hamarneh. "Star shape prior in fully convolutional networks for skin lesion segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 737–745.

[222] Zahra Mirikharaji, Yiqi Yan, and Ghassan Hamarneh. "Learning to segment skin lesions from noisy annotations." In: *MICCAI Workshop on Domain Adaptation and Representation Transfer*. Springer. 2019, pp. 207–215.

[223] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. "Deep learning with mixed supervision for brain tumor segmentation." In: *Journal of Medical Imaging* 6.3 (2019), pp. 034002–034002.

[224] Volodymyr Mnih and Geoffrey E Hinton. “Learning to label aerial images from noisy data.” In: *Proceedings of the 29th International conference on machine learning (ICML-12)*. 2012, pp. 567–574.

[225] Ahmed W Moawad, Anastasia Janas, Ujjwal Baid, Divya Ramakrishnan, Rachit Saluja, Nader Ashraf, Nazanin Maleki, Leon Jekel, Nikolay Yordanov, Pascal Fehringer, et al. “The brain tumor segmentation-metastases (brats-mets) challenge 2023: Brain metastasis segmentation on pre-treatment mri.” In: *ArXiv* (2024), arXiv–2306.

[226] Philip Müller, Felix Meissen, Johannes Brandt, Georgios Kaissis, and Daniel Rueckert. “Anatomy-driven pathology detection on chest x-rays.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 57–66.

[227] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. “When does label smoothing help?” In: *Advances in neural information processing systems* 32 (2019).

[228] Keelin Murphy, Bram Van Ginneken, Joseph M Reinhardt, Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E Christensen, Vincent Garcia, et al. “Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge.” In: *IEEE transactions on medical imaging* 30.11 (2011), pp. 1901–1920.

[229] Gowtham Krishnan Murugesan, Diana McCrumb, Eric Brunner, Jithendra Kumar, Rahul Soni, Vasily Grigorash, Stephen Moore, and Jeff Van Oss. “Improving lesion segmentation in FDG-18 whole-body PET/CT scans using multilabel approach: Autopet II challenge.” In: *arXiv preprint arXiv:2311.01574* (2023).

[230] Andriy Myronenko. “3D MRI brain tumor segmentation using autoencoder regularization.” In: *International MICCAI brainlesion workshop*. Springer. 2018, pp. 311–320.

[231] Reabal Najjar. “Redefining radiology: a review of artificial intelligence integration in medical imaging.” In: *Diagnostics* 13.17 (2023), p. 2760.

[232] Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. “Vila-m3: Enhancing vision-language models with medical expert knowledge.” In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. 2025, pp. 14788–14798.

[233] Fernando Navarro, Suprosanna Shit, Ivan Ezhov, Johannes Paetzold, Andrei Gafita, Jan C Peeken, Stephanie E Combs, and Bjoern H Menze. “Shape-aware complementary-task learning for multi-organ segmentation.” In: *International workshop on machine learning in medical imaging*. Springer. 2019, pp. 620–627.

[234] Nahida Nazir, Abid Sarwar, and Baljit Singh Saini. “Recent developments in denoising medical images using deep learning: An overview of models, techniques, and challenges.” In: *Micron* 180 (2024), p. 103615.

[235] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernadino Romera-Paredes, et al. "Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study." In: *Journal of medical Internet research* 23.7 (2021), e26151.

[236] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy." In: *arXiv preprint arXiv:1809.04430* (2018).

[237] Guy Nir, Soheil Hor, Davood Karimi, Ladan Fazli, Brian F Skinnider, Peyman Tavassoli, Dmitry Turbin, Carlos F Villamil, Gang Wang, R Storey Wilson, et al. "Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts." In: *Medical image analysis* 50 (2018), pp. 167–180.

[238] Tufve Nyholm, Stina Svensson, Sebastian Andersson, Joakim Jonsson, Maja Sohlin, Christian Gustafsson, Elisabeth Kjellén, Karin Söderström, Per Albertsson, Lennart Blomqvist, et al. "MR and CT data with multiobserver delineations of organs in the pelvic area—Part of the Gold Atlas project." In: *Medical physics* 45.3 (2018), pp. 1295–1300.

[239] Ozan Oktay, Enzo Ferrante, Konstantinos Kamnitsas, Mattias Heinrich, Wenjia Bai, Jose Caballero, Stuart A Cook, Antonio De Marvao, Timothy Dawes, Declan P O'Regan, et al. "Anatomically constrained neural networks (ACNNs): application to cardiac image enhancement and segmentation." In: *IEEE transactions on medical imaging* 37.2 (2017), pp. 384–395.

[240] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. "Attention u-net: Learning where to look for the pancreas." In: *arXiv preprint arXiv:1804.03999* (2018).

[241] Valentin Oreiller, Vincent Andrearczyk, Mario Jreige, Sarah Boughdad, Hesham Elhalawani, Joel Castelli, Martin Vallieres, Simeng Zhu, Juanying Xie, Ying Peng, et al. "Head and neck tumor segmentation in PET/CT: the HECKTOR challenge." In: *Medical image analysis* 77 (2022), p. 102336.

[242] Nobuyuki Otsu et al. "A threshold selection method from gray-level histograms." In: *Automatica* 11.285-296 (1975), pp. 23–27.

[243] Peiling Ou, Ru Wen, Lihua Deng, Linfeng Shi, Hongqin Liang, Jian Wang, and Chen Liu. "Exploring the changing landscape of medical imaging: Insights from highly cited studies before and during the COVID-19 pandemic." In: *European Radiology* 35.5 (2025), pp. 2922–2931.

[244] Nikos Paragios. “A level set approach for shape-driven segmentation and tracking of the left ventricle.” In: *IEEE transactions on medical imaging* 22.6 (2003), pp. 773–776.

[245] Junghoan Park, Sungeun Park, Han-Jae Chung, Da In Lee, Jong-min Kim, Se Hyung Kim, Eun Kyung Choe, Kyu Joo Park, and Soon Ho Yoon. “Deep learning for automatic volumetric bowel segmentation on body CT images.” In: *European Radiology* (2025), pp. 1–13.

[246] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. “Human–machine partnership with artificial intelligence for chest radiograph diagnosis.” In: *NPJ digital medicine* 2.1 (2019), p. 111.

[247] Jizong Peng, Guillermo Estrada, Marco Pedersoli, and Christian Desrosiers. “Deep co-training for semi-supervised image segmentation.” In: *Pattern Recognition* 107 (2020), p. 107269.

[248] Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. “Regularizing neural networks by penalizing confident output distributions.” In: *arXiv preprint arXiv:1701.06548* (2017).

[249] Hieu H Pham, Tung T Le, Dat Q Tran, Dat T Ngo, and Ha Q Nguyen. “Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels.” In: *Neurocomputing* 437 (2021), pp. 186–194.

[250] S. Pieper et al. *Spine metastatic bone cancer: pre and post radiotherapy CT (Spine-Mets-CT-SEG) [Dataset]*. Version Version 1. 2024. DOI: `10.7937/kh36-ds04`. URL: `https://doi.org/10.7937/kh36-ds04`.

[251] Walter HL Pinaya, Petru-Daniel Tudosiu, Robert Gray, Geraint Rees, Parashkev Nachev, Sebastien Ourselin, and M Jorge Cardoso. “Unsupervised brain imaging 3D anomaly detection and segmentation with transformers.” In: *Medical Image Analysis* 79 (2022), p. 102475.

[252] Passin Pornvoraphat, Kasenee Tiankanon, Rapat Pittayanon, Phanukorn Sunthornwetchapong, Peerapon Vateekul, and Rungsun Rerknimitr. “Real-time gastric intestinal metaplasia diagnosis tailored for bias and noisy-labeled data with multiple endoscopic imaging.” In: *Computers in Biology and Medicine* 154 (2023), p. 106582.

[253] Dandan Qiu, Jianguo Ju, Shumin Ren, Tongtong Zhang, Huijuan Tu, Xin Tan, and Fei Xie. “A deep learning-based cascade algorithm for pancreatic tumor segmentation.” In: *Frontiers in Oncology* 14 (2024), p. 1328146.

[254] Zhanhong Qiu, Haitao Gan, Ming Shi, Zhongwei Huang, and Zhi Yang. “Self-training with dual uncertainty for semi-supervised MRI image segmentation.” In: *Biomedical Signal Processing and Control* 94 (2024), p. 106355.

[255] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. "Annotating 8,000 Abdominal CT Volumes for Multi-Organ Segmentation in Three Weeks." In: *arXiv preprint arXiv:2305.09666* (2023).

[256] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Yucheng Tang, Alan L Yuille, Zongwei Zhou, et al. "Abdomenatlas-8k: Annotating 8,000 CT volumes for multi-organ segmentation in three weeks." In: *Advances in Neural Information Processing Systems* 36 (2023).

[257] Martin Rajchl, Matthew CH Lee, Ozan Oktay, Konstantinos Kamnitsas, Jonathan Passerat-Palmbach, Wenjia Bai, Mellisa Damodaram, Mary A Rutherford, Joseph V Hajnal, Bernhard Kainz, et al. "Deepcut: Object segmentation from bounding box annotations using convolutional neural networks." In: *IEEE transactions on medical imaging* 36.2 (2016), pp. 674–683.

[258] Dhanush Babu Ramesh, Rishika Iytha Sridhar, Pulakesh Upadhyaya, and Rishikesan Kamaleswaran. "Lung grounded-SAM (LuGSAM): A novel framework for integrating text prompts to segment anything model (SAM) for segmentation tasks of ICU chest X-Rays." In: *Authorea Preprints* (2023).

[259] Patrik F Raudaschl, Paolo Zaffino, Gregory C Sharp, Maria Francesca Spadea, Antong Chen, Benoit M Dawant, Thomas Albrecht, Tobias Gass, Christoph Langguth, Marcel Lüthi, et al. "Evaluation of segmentation methods on head and neck CT: auto-segmentation challenge 2015." In: *Medical physics* 44.5 (2017), pp. 2020–2036.

[260] Shubhankar Rawat, KPS Rana, and Vineet Kumar. "A novel complex-valued convolutional neural network for medical image denoising." In: *Biomedical Signal Processing and Control* 69 (2021), p. 102859.

[261] Faisal Rehman, Syed Irtiza Ali Shah, M Naveed Riaz, S Omer Gilani, and Faiza R. "A region-based deep level set formulation for vertebral bone segmentation of osteoporotic fractures." In: *Journal of digital imaging* 33.1 (2020), pp. 191–203.

[262] Annika Reinke, Minu D Tizabi, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, A Emre Kavur, Tim Rädsch, Carole H Sudre, Laura Acion, Michela Antonelli, et al. "Understanding metric-related pitfalls in image analysis validation." In: *Nature methods* 21.2 (2024), pp. 182–194.

[263] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. "Decoupled Semantic Prototypes enable learning from arbitrary annotation types for semi-weakly segmentation in expert-driven domains." In: *Accepted to IEEE/CVF conference on computer vision and pattern recognition*. 2023.

[264] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. "Grounded sam: Assembling open-world models for diverse visual tasks." In: *arXiv preprint arXiv:2401.14159* (2024).

[265] Blaine Rister, Darvin Yi, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. "CT-ORG, a new dataset for multiple organ segmentation in computed tomography." In: *Scientific Data* 7.1 (2020), p. 381.

[266] Maximilian Rokuss, Balint Kovacs, Yannick Kirchhoff, Shuhan Xiao, Constantin Ulrich, Klaus H Maier-Hein, and Fabian Isensee. "From FDG to PSMA: A Hitchhiker's Guide to Multitracer, Multicenter Lesion Segmentation in PET/CT Imaging." In: *arXiv preprint arXiv:2409.09478* (2024).

[267] Yi Rong, Quan Chen, Yabo Fu, Xiaofeng Yang, Hania A Al-Hallaq, Q Jackie Wu, Lulin Yuan, Ying Xiao, Bin Cai, Kujtim Latifi, et al. "NRG oncology assessment of artificial intelligence deep learning–based auto-segmentation for radiation therapy: current developments, clinical considerations, and future directions." In: *International Journal of Radiation Oncology* Biology* Physics* 119.1 (2024), pp. 261–280.

[268] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[269] Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. "Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation." In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part I 18*. Springer. 2015, pp. 556–564.

[270] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. "" GrabCut" interactive foreground extraction using iterated graph cuts." In: *ACM transactions on graphics (TOG)* 23.3 (2004), pp. 309–314.

[271] Pierre Rougé, Odyssée Merveille, and Nicolas Passat. "ccDice: A topology-aware Dice score based on connected components." In: *International Workshop on Topology-and Graph-Informed Imaging Informatics*. Springer. 2024, pp. 11–21.

[272] Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F Jaeger, and Klaus H Maier-Hein. "Mednext: transformer-driven scaling of convnets for medical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 405–415.

[273] Saikat Roy, Tassilo Wald, Gregor Koehler, Maximilian R. Rokuss, Nico Disch, Julius Holzschuh, David Zimmerer, and Klaus H. Maier-Hein. *SAM.MD: Zero-shot medical image segmentation capabilities of the Segment Anything Model*. en. arXiv:2304.05396 [cs, eess]. Apr. 2023. URL: http://arxiv.org/abs/2304.05396 (visited on 04/29/2023).

[274] Rina D Rudyanto, Sjoerd Kerkstra, Eva M Van Rikxoort, Catalin Fetita, Pierre-Yves Brillet, Christophe Lefevre, Wenzhe Xue, Xiangjun Zhu, Jianming Liang, Ilkay Öksüz, et al. "Comparing algorithms for automated vessel segmentation in computed tomography scans of the lung: the VESSEL12 study." In: *Medical image analysis* 18.7 (2014), pp. 1217–1232.

[275] Sameera V Mohd Sagheer and Sudhish N George. "A review on medical image denoising algorithms." In: *Biomedical signal processing and control* 61 (2020), p. 102036.

[276] Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J Erickson. "Interactive segmentation of medical images through fully convolutional neural networks." In: *arXiv preprint arXiv:1903.08205* (2019).

[277] Claudio E von Schacky, Nikolas J Wilhelm, Valerie S Schäfer, Yannik Leonhardt, Felix G Gassert, Sarah C Foreman, Florian T Gassert, Matthias Jung, Pia M Jungmann, Maximilian F Russe, et al. "Multitask deep learning for segmentation and classification of primary bone tumors on radiographs." In: *Radiology* 301.2 (2021), pp. 398–406.

[278] Arne Schmidt, Pablo Morales-Alvarez, and Rafael Molina. "Probabilistic modeling of inter-and intra-observer variability in medical image segmentation." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2023, pp. 21097–21106.

[279] Manuel Schultheiss, Philipp Schmette, Jannis Bodden, Juliane Aichele, Christina Müller-Leisse, Felix G Gassert, Florian T Gassert, Joshua F Gawlitza, Felix C Hofmann, Daniel Sasse, et al. "Lung nodule detection in chest X-rays using synthetic ground-truth data comparing CNN-based diagnosis to human performance." In: *Scientific Reports* 11.1 (2021), p. 15857.

[280] Constantin Seibold, Alexander Jaus, Matthias A Fink, Moon Kim, Simon Reiß, Ken Herrmann, Jens Kleesiek, and Rainer Stiefelhagen. "Accurate fine-grained segmentation of human anatomy in radiographs via volumetric pseudo-labeling." In: *arXiv preprint arXiv:2306.03934* (2023).

[281] Silvia Seidlitz, Jan Sellner, Jan Odenthal, Berkin Özdemir, Alexander Studier-Fischer, Samuel Knödler, Leonardo Ayala, Tim J Adler, Hannes G Kenngott, Minu Tizabi, et al. "Robust deep learning-based semantic organ segmentation in hyperspectral images." In: *Medical Image Analysis* 80 (2022), p. 102488.

[282] Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. "VerSe: a vertebrae labelling and segmentation benchmark for multi-detector CT images." In: *Medical image analysis* 73 (2021), p. 102166.

[283] Raghavendra Selvan, Thomas Kipf, Max Welling, Antonio Garcia-Uceda Juarez, Jesper H Pedersen, Jens Petersen, and Marleen de Bruijne. “Graph refinement based airway extraction using mean-field networks and graph neural networks.” In: *Medical image analysis* 64 (2020), p. 101751.

[284] Raghavendra Selvan, Thomas Kipf, Max Welling, Jesper H Pedersen, Jens Petersen, and Marleen de Bruijne. “Extraction of airways using graph neural networks.” In: *arXiv preprint arXiv:1804.04436* (2018).

[285] Yiqing Shen, Jingxing Li, Xinyuan Shao, Blanca Inigo Romillo, Ankush Jindal, David Dreizin, and Mathias Unberath. “Fastsam3d: An efficient segment anything model for 3d volumetric medical images.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 542–552.

[286] Jialin Shi, Kailai Zhang, Chenyi Guo, Youquan Yang, Yali Xu, and Ji Wu. “A survey of label-noise deep learning for medical image analysis.” In: *Medical image analysis* 95 (2024), p. 103166.

[287] Xue Shi and Chunming Li. “Anatomical knowledge based level set segmentation of cardiac ventricles from MRI.” In: *Magnetic Resonance Imaging* 86 (2022), pp. 135–148.

[288] Seung Yeon Shin, Soochahn Lee, Il Dong Yun, and Kyoung Mu Lee. “Deep vessel segmentation by learning graphical connectivity.” In: *Medical image analysis* 58 (2019), p. 101556.

[289] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. “Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists’ detection of pulmonary nodules.” In: *American journal of roentgenology* 174.1 (2000), pp. 71–74.

[290] Rushin Shojaii, Javad Alirezaie, and Paul Babyn. “Automatic lung segmentation in CT images using watershed transform.” In: *IEEE international conference on image processing 2005*. Vol. 2. IEEE. 2005, pp. II–1270.

[291] Giovanni LF da Silva, Petterson S Diniz, Jonnison L Ferreira, Joao VF Franca, Aristofanes C Silva, Anselmo C de Paiva, and Elton AA de Cavalcanti. “Superpixel-based deep convolutional neural networks and active contour model for automatic prostate segmentation on 3D MRI scans.” In: *Medical & Biological Engineering & Computing* 58.9 (2020), pp. 1947–1964.

[292] William Silversmith. *cc3d: Connected components on multilabel 3D & 2D images*. 2021.

[293] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition.” In: *arXiv preprint arXiv:1409.1556* (2014).

[294] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. “A large annotated medical image dataset for the development and evaluation of segmentation algorithms.” In: *arXiv preprint arXiv:1902.09063* (2019).

[295] Luc Soler, Alexandre Hostettler, Vincent Agnus, Arnaud Charnoz, J Fasquel, Johan Moreau, A Osswald, Mourad Bouhadjar, and Jacques Marescaux. “3D image reconstruction for comparison of algorithm database: A patient specific anatomical and medical image database.” In: *IRCAD, Strasbourg, France, Tech. Rep* 1.1 (2010).

[296] Eva Maria Stoiber et al. “Analyzing human decisions in IGRT of head-and-neck cancer patients to teach image registration algorithms what experts know.” In: *Radiation Oncology* 12 (2017), pp. 1–7.

[297] Martin Styner, Joohwi Lee, Brian Chin, M Chin, Olivier Commowick, H Tran, Silva Markovic-Plese, Valerie Jewells, and Simon Warfield. “3D segmentation in the clinic: A grand challenge II: MS lesion segmentation.” In: *MIDAS journal* 2008 (2008), pp. 1–6.

[298] Carole H Sudre, Beatriz Gomez Anson, Silvia Ingala, Chris D Lane, Daniel Jimenez, Lukas Haider, Thomas Varsavsky, Ryutaro Tanno, Lorna Smith, Sébastien Ourselin, et al. “Let’s agree to disagree: Learning highly debatable multirater labelling.” In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 665–673.

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

[300] Lalith Kumar Shiyam Sundar, Josef Yu, Otto Muzik, Oana C Kulterer, Barbara Fueger, Daria Kifjak, Thomas Nakuz, Hyung Min Shin, Annika Katharina Sima, Daniela Kitzmantl, et al. “Fully automated, semantic segmentation of whole-body 18F-FDG PET/CT images based on data-centric artificial intelligence.” In: *Journal of Nuclear Medicine* 63.12 (2022), pp. 1941–1948.

[301] Abdel Aziz Taha and Allan Hanbury. “Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool.” In: *BMC Medical Imaging* 15.1 (2015), p. 29. DOI: 10.1186/s12880-015-0068-x. URL: https://doi.org/10.1186/s12880-015-0068-x.

[302] Hima Tallam, Daniel C Elton, Sungwon Lee, Paul Wakim, Perry J Pickhardt, and Ronald M Summers. “Fully automated abdominal CT biomarkers for type 2 diabetes using deep learning.” In: *Radiology* 304.1 (2022), pp. 85–95.

[303] Min Tang, Sepehr Valipour, Zichen Zhang, Dana Cobzas, and Martin Jagersand. "A deep level set method for image segmentation." In: *International Workshop on Deep Learning in Medical Image Analysis*. Springer. 2017, pp. 126–134.

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

[305] Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C Alexander, and Nathan Silberman. "Learning from noisy labels by regularized estimation of annotator confusion." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019, pp. 11244–11253.

[306] The Royal College of Radiologists. *Clinical Radiology Workforce Census 2023*. Technical Report. Data as of Oktober 2023. London, UK: The Royal College of Radiologists, June 2024. URL: https://www.rcr.ac.uk/media/5befglss/rcr-census-clinical-radiology-workforce-census-2023.pdf (visited on 06/09/2025).

[307] Bethany H Thompson, Gaetano Di Caterina, and Jeremy P Voisey. "Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation." In: *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE. 2022, pp. 1–5.

[308] Therese Tillin, Nita G Forouhi, Paul M McKeigue, and Nish Chaturvedi. "Southall And Brent REvisited: Cohort profile of SABRE, a UK population-based comparison of cardiovascular disease and diabetes in people of European, Indian Asian and African Caribbean origins." In: *International journal of epidemiology* 41.1 (2012), pp. 33–42.

[309] Constantin Ulrich, Fabian Isensee, Tassilo Wald, Maximilian Zenk, Michael Baumgartner, and Klaus H Maier-Hein. "Multitalent: A multi-dataset approach to medical image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2023, pp. 648–658.

[310] Bram Van Ginneken, Tobias Heimann, and Martin Styner. "3D segmentation in the clinic: A grand challenge." In: *MICCAI workshop on 3D segmentation in the clinic: a grand challenge*. Vol. 1. 2007, pp. 7–15.

[311] C. J. Van Rijsbergen. *Information Retrieval*. Newton, MA, USA: Butterworth-Heinemann, 1979.

[312] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." In: *Advances in neural information processing systems* 30 (2017).

[313] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. "SciPy 1.0: fundamental algorithms for scientific computing in Python." In: *Nature methods* 17.3 (2020), pp. 261–272.

[314] Alexandra Walter, Cornelius J Bauer, Ama Katseena Yawson, Philipp Hoegen-Saßmannshausen, Sebastian Adeberg, Jürgen Debus, Oliver Jäkel, Martin Frank, and Kristina Giske. "Accuracy of an articulated head-and-neck motion model using deep learning-based instance segmentation of skeletal bones in CT scans for image registration in radiotherapy." In: *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* 13.1 (2025), p. 2455752.

[315] Alexandra Walter, Philipp Hoegen-Saßmannshausen, Goran Stanic, Joao Pedro Rodrigues, Sebastian Adeberg, Oliver Jäkel, Martin Frank, and Kristina Giske. "Segmentation of 71 anatomical structures necessary for the evaluation of guideline-conforming clinical target volumes in head and neck cancers." In: *Cancers* 16.2 (2024).

[316] Guotai Wang, Xinglong Liu, Chaoping Li, Zhiyong Xu, Jiugen Ruan, Haifeng Zhu, Tao Meng, Kang Li, Ning Huang, and Shaoting Zhang. "A noise-robust framework for automatic segmentation of COVID-19 pneumonia lesions from CT images." In: *IEEE Transactions on Medical Imaging* 39.8 (2020), pp. 2653–2663.

[317] Haonan Wang and Xiaomeng Li. "Dhc: Dual-debiased heterogeneous co-training framework for class-imbalanced semi-supervised medical image segmentation." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2023, pp. 582–591.

[318] Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, et al. "Sam-med3d: towards general-purpose segmentation models for volumetric medical images." In: *European Conference on Computer Vision*. Springer. 2024, pp. 51–67.

[319] Hongzhi Wang, Jung W Suh, Sandhitsu R Das, John B Pluta, Caryne Craige, and Paul A Yushkevich. "Multi-atlas segmentation with joint label fusion." In: *IEEE transactions on pattern analysis and machine intelligence* 35.3 (2012), pp. 611–623.

[320] Kang Wang, Zeyang Li, Haoran Wang, Siyu Liu, Mingyuan Pan, Manning Wang, Shuo Wang, and Zhijian Song. "Improving brain tumor segmentation with anatomical prior-informed pre-training." In: *Frontiers in Medicine* 10 (2023), p. 1211800.

[321] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. "Omni-DETR: Omni-supervised object detection with transformers." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 9367–9376.

[322] Ping Wang, Jizong Peng, Marco Pedersoli, Yuanfeng Zhou, Caiming Zhang, and Christian Desrosiers. "Self-paced and self-consistent co-training for semi-supervised image segmentation." In: *Medical Image Analysis* 73 (2021), p. 102146.

[323] Weiyue Wang and Ulrich Neumann. "Depth-aware cnn for rgb-d segmentation." In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 135–150.

[324] Wenji Wang, Qing Xia, Zhiqiang Hu, Zhennan Yan, Zhuowei Li, Yang Wu, Ning Huang, Yue Gao, Dimitris Metaxas, and Shaoting Zhang. "Few-shot learning by a cascaded framework with shape-constrained pseudo label assessment for whole heart segmentation." In: *IEEE Transactions on Medical Imaging* 40.10 (2021), pp. 2629–2641.

[325] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. "Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 2097–2106.

[326] Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, Yuyin Zhou, Wei Shen, Elliot K Fishman, and Alan L Yuille. "Deep distance transform for tubular structure segmentation in ct scans." In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 3833–3842.

[327] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. "Symmetric cross entropy for robust learning with noisy labels." In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 322–330.

[328] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. "Cris: Clip-driven referring image segmentation." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 11686–11695.

[329] Zhepeng Wang, Runxue Bao, Yawen Wu, Guodong Liu, Lei Yang, Liang Zhan, Feng Zheng, Weiwen Jiang, and Yanfu Zhang. "Self-guided knowledge-injected graph neural network for alzheimer's diseases." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2024, pp. 378–388.

[330] Zhiwei Wang, Chaoyue Liu, Danpeng Cheng, Liang Wang, Xin Yang, and Kwang-Ting Cheng. "Automated detection of clinically significant prostate cancer in mp-MRI images based on an end-to-end deep neural network." In: *IEEE transactions on medical imaging* 37.5 (2018), pp. 1127–1139.

[331] Ziyang Wang and Irina Voiculescu. “Dealing with unreliable annotations: a noise-robust network for semantic segmentation through a transformer-improved encoder and convolution decoder.” In: *Applied Sciences* 13.13 (2023), p. 7966.

[332] Simon K Warfield, Kelly H Zou, and William M Wells. “Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation.” In: *IEEE transactions on medical imaging* 23.7 (2004), pp. 903–921.

[333] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. “TotalSegmentator: robust segmentation of 104 anatomic structures in CT images.” In: *Radiology: Artificial Intelligence* 5.5 (2023), e230024.

[334] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. “TotalSegmentator: robust segmentation of 104 anatomical structures in CT images.” In: (Aug. 2022). arXiv: 2208.05868. URL: http://arxiv.org/abs/2208.05868.

[335] Jakob Wasserthal, Manfred Meyer, Hanns-Christian Breit, Joshy Cyriac, Shan Yang, and Martin Segeroth. “TotalSegmentator: robust segmentation of 104 anatomical structures in CT images.” In: *arXiv preprint arXiv:2208.05868* (2022).

[336] Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. “Diffusion models for medical anomaly detection.” In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2022, pp. 35–45.

[337] Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. “Scribbleprompt: fast and flexible interactive segmentation for any biomedical image.” In: *European Conference on Computer Vision*. Springer. 2024, pp. 207–229.

[338] Hallee E. Wong, Marianne Rakic, John Guttag, and Adrian V. Dalca. “ScribblePrompt: Fast and Flexible Interactive Segmentation for Any Biomedical Image.” In: *European Conference on Computer Vision (ECCV)* (2024).

[339] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. “Cbam: Convolutional block attention module.” In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 3–19.

[340] Huimin Wu, Xiaomeng Li, Yiqun Lin, and Kwang-Ting Cheng. “Compete to win: Enhancing pseudo labels for barely-supervised medical image segmentation.” In: *IEEE Transactions on Medical Imaging* 42.11 (2023), pp. 3244–3255.

[341] Junde Wu, Ziyue Wang, Mingxuan Hong, Wei Ji, Huazhu Fu, Yanwu Xu, Min Xu, and Yueming Jin. “Medical sam adapter: Adapting segment anything model for medical image segmentation.” In: *Medical image analysis* 102 (2025), p. 103547.

[342] Yingda Xia, Fengze Liu, Dong Yang, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. "3d semi-supervised learning with uncertainty-aware multi-view co-training." In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 2020, pp. 3646–3655.

[343] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. "Sun database: Large-scale scene recognition from abbey to zoo." In: *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE. 2010, pp. 3485–3492.

[344] Li Xiao, Yinhao Li, Luxi Qv, Xinxia Tian, Yijie Peng, and S Kevin Zhou. "Pathological image segmentation with noisy labels." In: *arXiv preprint arXiv:2104.02602* (2021).

[345] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. "Self-training with noisy student improves imagenet classification." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2020, pp. 10687–10698.

[346] Moucheng Xu, Yukun Zhou, Chen Jin, Marius de Groot, Daniel C Alexander, Neil P Oxtoby, Yipeng Hu, and Joseph Jacob. "Expectation maximisation pseudo labels." In: *Medical Image Analysis* 94 (2024), p. 103125.

[347] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. "Deep interactive object selection." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 373–381.

[348] Zhenlin Xu and Marc Niethammer. "DeepAtlas: Joint semi-supervised learning of image registration and segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 420–429.

[349] Zhoubing Xu, Christopher P Lee, Mattias P Heinrich, Marc Modat, Daniel Rueckert, Sebastien Ourselin, Richard G Abramson, and Bennett A Landman. "Evaluation of six registration methods for the human abdomen on clinically acquired CT." In: *IEEE Transactions on Biomedical Engineering* 63.8 (2016), pp. 1563–1572.

[350] Cheng Xue, Qiao Deng, Xiaomeng Li, Qi Dou, and Pheng-Ann Heng. "Cascaded robust learning at imperfect labels for chest x-ray segmentation." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2020, pp. 579–588.

[351] Cheng Xue, Qi Dou, Xueying Shi, Hao Chen, and Pheng-Ann Heng. "Robust learning at noisy labeled medical images: Applied to skin lesion classification." In: *2019 IEEE 16th International symposium on biomedical imaging (ISBI 2019)*. IEEE. 2019, pp. 1280–1283.

[352] Ke Yan, Youbao Tang, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M Summers. "MULAN: multitask universal lesion analysis network for joint lesion detection, tagging, and segmentation." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2019, pp. 194–202.

[353] Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning." In: *Journal of medical imaging* 5.3 (2018), pp. 036501–036501.

[354] Jiancheng Yang, Shixuan Gu, Donglai Wei, Hanspeter Pfister, and Bingbing Ni. "RibSeg Dataset and Strong Point Cloud Baselines for Rib Segmentation from CT Scans." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2021, pp. 611–621.

[355] Jiancheng Yang, Shixuan Gu, Donglai Wei, Hanspeter Pfister, and Bingbing Ni. "Ribseg dataset and strong point cloud baselines for rib segmentation from ct scans." In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2021, pp. 611–621.

[356] Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. "The limits of fair medical imaging AI in real-world generalization." In: *Nature Medicine* 30.10 (2024), pp. 2838–2848.

[357] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. "Lavt: Language-aware vision transformer for referring image segmentation." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 18155–18165.

[358] Jianhua Yao, Joseph E Burns, Hector Munoz, and Ronald M Summers. "Detection of vertebral body fractures based on cortical shell unwrapping." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2012, pp. 509–516.

[359] Shaohan Yin, Xiao Luo, Yadi Yang, Ying Shao, Lidi Ma, Cuiping Lin, Qiuxia Yang, Deling Wang, Yingwei Luo, Zhijun Mai, et al. "Development and validation of a deep-learning model for detecting brain metastases on 3D post-contrast MRI: a multi-center multi-reader evaluation study." In: *Neuro-oncology* 24.9 (2022), pp. 1559–1570.

[360] Shaode Yu, Mingli Chen, Erlei Zhang, Junjie Wu, Hang Yu, Zi Yang, Lin Ma, Xuejun Gu, and Weiguo Lu. "Robustness study of noisy annotation in deep learning based medical image segmentation." In: *Physics in Medicine & Biology* 65.17 (2020), p. 175007.

[361] Qian Yue, Xinzhe Luo, Qing Ye, Lingchao Xu, and Xiahai Zhuang. "Cardiac segmentation from LGE MRI using deep neural network incorporating shape and spatial priors." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 559–567.

[362] Fatemeh Zabihollahy, James A White, and Eranga Ukwatta. "Convolutional neural network-based approach for segmentation of left ventricle myocardial scar from 3D late gadolinium enhancement MR images." In: *Medical physics* 46.4 (2019), pp. 1740–1751.

[363] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. "Open-vocabulary object detection using captions." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 14393–14402.

[364] Hejia Zhang, Xia Zhu, and Theodore L Willke. "Segmenting brain tumors with symmetry." In: *arXiv preprint arXiv:1711.06636* (2017).

[365] Ju Zhang, Weiwei Gong, Lieli Ye, Fanghong Wang, Zhibo Shangguan, and Yun Cheng. "A review of deep learning methods for denoising of medical low-dose CT images." In: *Computers in Biology and Medicine* 171 (2024), p. 108112.

[366] Le Zhang, Ryutaro Tanno, Mou-Cheng Xu, Chen Jin, Joseph Jacob, Olga Ciccarelli, Frederik Barkhof, and Daniel Alexander. "Disentangling human error from ground truth in segmentation of medical images." In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15750–15762.

[367] Minghui Zhang, Yangqian Wu, Hanxiao Zhang, Yulei Qin, Hao Zheng, Wen Tang, Corey Arnold, Chenhao Pei, Pengxin Yu, Yang Nan, et al. "Multi-site, Multi-domain Airway Tree Modeling." In: *Medical Image Analysis* 90 (2023), p. 102957.

[368] Rongzhao Zhang, Zhian Bai, Ruoying Yu, Wenrao Pang, Lingyun Wang, Lifeng Zhu, Xiaofan Zhang, Huan Zhang, and Weiguo Hu. "Ag-crc: Anatomy-guided colorectal cancer segmentation in ct with imperfect anatomical knowledge." In: *arXiv preprint arXiv:2310.04677* (2023).

[369] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. "Road extraction by deep residual u-net." In: *IEEE Geoscience and Remote Sensing Letters* 15.5 (2018), pp. 749–753.

[370] Tianyi Zhao, Kai Cao, Jiawen Yao, Isabella Nogues, Le Lu, Lingyun Huang, Jing Xiao, Zhaozheng Yin, and Ling Zhang. "3D graph anatomy geometry-integrated network for pancreatic mass segmentation, diagnosis, and quantitative patient management." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 13743–13752.

[371] Xiaomei Zhao, Yihong Wu, Guidong Song, Zhenye Li, Yazhuo Zhang, and Yong Fan. "A deep learning model integrating FCNNs and CRFs for brain tumor segmentation." In: *Medical image analysis* 43 (2018), pp. 98–111.

[372] Yu Zhao, Hongwei Li, Shaohua Wan, Anjany Sekuboyina, Xiaobin Hu, Giles Tetteh, Marie Piraud, and Bjoern Menze. "Knowledge-aided convolutional neural network for small organ segmentation." In: *IEEE journal of biomedical and health informatics* 23.4 (2019), pp. 1363–1373.

[373] Hao Zheng, Susan M Motch Perrine, M Kathleen Pitirri, Kazuhiko Kawasaki, Chaoli Wang, Joan T Richtsmeier, and Danny Z Chen. "Cartilage segmentation in high-resolution 3D micro-CT images via uncertainty-guided self-training with very sparse annotation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2020, pp. 802–812.

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

[375] Tao Zhou, Xinyu Ye, Huiling Lu, Yujie Guo, Hongxia Wang, and Yang Liu. "An adaptive and lightweight YOLOv5 detection model for lung tumor in PET/CT images." In: *Scientific Reports* 14.1 (2024), p. 30719.

[376] Yuyin Zhou, Yan Wang, Peng Tang, Song Bai, Wei Shen, Elliot Fishman, and Alan Yuille. "Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training." In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2019, pp. 121–140.

[377] Zijian Zhou, Oluwatosin Alabi, Meng Wei, Tom Vercauteren, and Miaojing Shi. "Text promptable surgical instrument segmentation with vision-language models." In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 28611–28623.

[378] Haidong Zhu, Jialin Shi, and Ji Wu. "Pick-and-learn: Automatic quality evaluation for noisy-labeled image segmentation." In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 576–584.

[379] Xiaojin Zhu and Andrew Goldberg. *Introduction to semi-supervised learning*. Morgan & Claypool Publishers, 2009.

[380] Xiaojin Jerry Zhu. "Semi-supervised learning literature survey." In: (2005).

[381] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. "Deformable detr: Deformable transformers for end-to-end object detection." In: *arXiv preprint arXiv:2010.04159* (2020).

[382] Mingrui Zhuang, Zhonghua Chen, Yuxin Yang, Lauri Kettunen, and Hongkai Wang. "Annotation-efficient training of medical image segmentation network based on scribble guidance in difficult areas." In: *International Journal of Computer Assisted Radiology and Surgery* 19.1 (2024), pp. 87–96.

[383] Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, et al. "Evaluation of algorithms for multi-modality whole heart segmentation: an open-access grand challenge." In: *Medical image analysis* 58 (2019), p. 101537.

[384] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. "Generalized decoding for pixel, image, and language." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2023, pp. 15116–15127.

[385] Maria A Zuluaga, M Jorge Cardoso, Marc Modat, and Sébastien Ourselin. "Multi-atlas propagation whole heart segmentation from MRI and CTA using a local normalised correlation coefficient criterion." In: *International Conference on Functional Imaging and Modeling of the Heart*. Springer. 2013, pp. 174–181.

[386] Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. "Medxpertqa: Benchmarking expert-level medical reasoning and understanding." In: *arXiv preprint arXiv:2501.18362* (2025).