Karlsruher Institut für Technologie

SEMI-WEAKLY SUPERVISED LEARNING FOR LABEL-EFFICIENT SEMANTIC SEGMENTATION IN EXPERT-DRIVEN DOMAINS

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
SIMON MICHAEL REIß
aus Backnang

Tag der mündlichen Prüfung: 06.11.2023
Hauptreferent: Prof. Dr.-Ing. Rainer Stiefelhagen
Korreferent: Prof. Dr. Zeynep Akata

Abstract

Through deep learning, semantic segmentation systems have been utilized to yield impressive results, yet this was achieved on the basis of supervised learning which is limited by the availability of costly, pixel-wise annotated images. When investigating the performance of these segmentation systems in contexts where annotations are scarce they fall short of the high expectations induced by their performance in annotation rich scenarios. This predicament weights especially heavy, when the annotations have to be provided by heavily trained personnel, *e.g.* medical doctors, process experts or scientists. To bring well-performing segmentation models into these annotation deprived expert-driven domains, new solutions are needed.

To this intent, we first investigate how badly current segmentation models really cope with extreme annotation scarce settings in expert-driven imagery domains. This is directly linked to the question whether costly pixel-wise annotations that high performing segmentation models are trained with can be circumvented, or if they are, conversely, a cost-effective kick-start to bring the segmentation off the ground when used sparingly. We further briefly dive into the question whether different kinds of annotations, weak- and full pixel-wise annotations with varying costs, can be used jointly in training segmentation systems in order to make the annotation process more flexible.

Expert-driven domains do not only come with annotation scarcity but with entirely different imaging properties, including a volumetric shape. Moving from 2D- to 3D semantic segmentation and training models in a supervised fashion entails voxel-wise annotation processes, which multiplies the time-expenditure for annotation with

the added dimension. To circumvent this costly process and end up at manageable ways of annotation, we investigate segmentation model training strategies which only require either more economical sparse annotations or unlabeled volumes. This shift in supervision type can bring annotation costs down for volumetric segmentation tasks and enable applications to be built in expert-driven domains. As side-effect annotators are freed from the laborious task of densely annotating entire volumes which reduces redundant work to be done due to visually redundant regions as present in volume data.

Finally, we ask the question, whether it is possible to free up expert annotators from the strict requirement of having to supply a single, specific annotation type and design a training strategy which can work with a broad diversity of semantic cues. We design a training strategy for this scenario and, in our extensive experimental evaluation, bring to light interesting properties of different annotation type mixes in relation to their resulting segmentation performance.

Our investigations led to new research directions in semi-weakly supervised segmentation, novel, annotation-efficient methods and training strategies as well as experimental insights which are valuable for practitioners and can improve annotation processes by making them annotation-efficient, expert-centric and flexible.

Zusammenfassung

Unter Zuhilfenahme von Deep Learning haben semantische Segmentierungssysteme beeindruckende Ergebnisse erzielt, allerdings auf der Grundlage von überwachtem Lernen, das durch die Verfügbarkeit kostspieliger, pixelweise annotierter Bilder limitiert ist. Bei der Untersuchung der Performance dieser Segmentierungssysteme in Kontexten, in denen kaum Annotationen vorhanden sind, bleiben sie hinter den hohen Erwartungen, die durch die Performance in annotationsreichen Szenarien geschürt werden, zurück. Dieses Dilemma wiegt besonders schwer, wenn die Annotationen von lange geschultem Personal, z.B. Medizinern, Prozessexperten oder Wissenschaftlern, erstellt werden müssen. Um gut funktionierende Segmentierungsmodelle in diese annotationsarmen, Experten-angetriebenen Domänen zu bringen, sind neue Lösungen nötig.

Zu diesem Zweck untersuchen wir zunächst, wie schlecht aktuelle Segmentierungsmodelle mit extrem annotationsarmen Szenarien in Experten-angetriebenen Bildgebungsdomänen zurechtkommen. Daran schließt sich direkt die Frage an, ob die kostspielige pixelweise Annotation, mit der Segmentierungsmodelle in der Regel trainiert werden, gänzlich umgangen werden kann, oder ob sie umgekehrt ein Kosten-effektiver Anstoß sein kann, um die Segmentierung in Gang zu bringen, wenn sie sparsam eingestetzt wird. Danach gehen wir auf die Frage ein, ob verschiedene Arten von Annotationen, schwache- und pixelweise Annotationen mit unterschiedlich hohen Kosten, gemeinsam genutzt werden können, um den Annotationsprozess flexibler zu gestalten.

Experten-angetriebene Domänen haben oft nicht nur einen Annotationsmangel, sondern auch völlig andere Bildeigenschaften, beispielsweise volumetrische Bild-Daten.

Der Übergang von der 2D- zur 3D-semantischen Segmentierung führt zu voxelweisen Annotationsprozessen, was den nötigen Zeitaufwand für die Annotierung mit der zusätzlichen Dimension multipliziert. Um zu einer handlicheren Annotation zu gelangen, untersuchen wir Trainingsstrategien für Segmentierungsmodelle, die nur preiswertere, partielle Annotationen oder rohe, nicht annotierte Volumina benötigen. Dieser Wechsel in der Art der Überwachung im Training macht die Anwendung der Volumensegmentierung in Experten-angetriebenen Domänen realistischer, da die Annotationskosten drastisch gesenkt werden und die Annotatoren von Volumina-Annotationen befreit werden, welche naturgemäß auch eine Menge visuell redundanter Regionen enthalten würden.
Schließlich stellen wir die Frage, ob es möglich ist, die Annotations-Experten von der strikten Anforderung zu befreien, einen einzigen, spezifischen Annotationstyp liefern zu müssen, und eine Trainingsstrategie zu entwickeln, die mit einer breiten Vielfalt semantischer Information funktioniert. Eine solche Methode wurde hierzu entwickelt und in unserer umfangreichen experimentellen Evaluierung kommen interessante Eigenschaften verschiedener Annotationstypen-Mixe in Bezug auf deren Segmentierungsperformance ans Licht.
Unsere Untersuchungen führten zu neuen Forschungsrichtungen in der semi-weakly überwachten Segmentierung, zu neuartigen, annotationseffizienteren Methoden und Trainingsstrategien sowie zu experimentellen Erkenntnissen, zur Verbesserung von Annotationsprozessen, indem diese annotationseffizient, expertenzentriert und flexibel gestaltet werden.

Acknowledgements

First off, I want to thank you Rainer, for taking a chance on me to pursue this wonderful research endeavour, for giving me guidance when I needed it and for letting me roam free to try out things and explore ideas which has been an exciting and stimulating challenge for my creativity. My thanks also extend to you Professor Akata, as you kindly agreed to serve as second reviewer on my committee, which makes me especially happy, as I found a lot of inspiration in your work which made me excited about rigorously exploring semi-weakly supervised learning. Alex, I also take inspiration in your positive, kind and always welcoming attitude, it was a delight pushing the limits with you and I am very grateful for the atmosphere and collaborative environment which you created – perfect for discussing ideas and learning from your immense technical knowledge and grasp of the field. Similarly, Erik, your enthusiasm for deeply understanding and for precisely, mathematically note down technical details gave me the ability to supply my scientific ideas with a voice to be understood, thank you so much for your lasting support even beyond your time at ZEISS. Be it small chats between the door and hinge or helping with complicated university procedures, Corinna, you made me feel welcome from when I first entered the lab. The first time, I peaked into the realm of hands-on Computer Vision as a student research assistant was for you, Alina, thank you for closely involving me in your research and helping me off the ground to bring my own ideas to flourishing, I owe a lot to you. Constantin, I did not just share an office with you, I went on an intense, creative journey with you, which we both did not know where it would lead us. Thank you for being such a great companion, now I know how the figure of

speech “the real treasure was the friends that we made along the way” was meant before it became a meme. And thanks for nudging me to clear my head outside of the lab, for letting me distract you from work with stupid ideas and for working side-by-side late into the night. Without anyone around to discuss and share thoughts with, this deeply social activity of research would be utterly depressing. Therefore, I want to sincerely thank you, Kailun, Marios, Manel, Monica, Tobias, Vivek, Jiaming, Saquib, David, Kunyu, Alex, Zdravko, Omar, Junwei, Yufan, Ruiping, it’s you all who created a kind, productive, thought provoking and cheerful environment of which I was fortunate enough to be a small part of. During my visits at ZEISS I was always greeted with kindness, openness and sincere interest as well as a contagious spirit of striving for actively seeking out the best solutions at the edge of what’s possible, thank you wonderful people at CRT for sharpening my view for what matters when transferring research into actual products. This collaboration has been a beacon of unconstrained and fruitful development and exchange of ideas, for which I am grateful to Tanja and all those at ZEISS who made it come to life. The bright students roaming over the campus shape the university environment. I want to thank three of them in particular, Paul, Johannes and Dima, with whom I had the pleasure to probe into exciting research directions. Starting a PhD at the same time as a world wide pandemic hits certainly comes with challenges, but all these challenges are bearable with the social web of such wonderful, long lasting friendships which I am immensely thankful to have. Domy – with your example the thought started to form: might it be possible for me too? Thank you for being a role model and for being there and being awesome pretty much my entire life!, Jan – ~~I still think particle physics is useless~~ thanks for spontaneously being up for anything, always, no questions asked, Nico – remember the summer we went skateboarding? I certainly do!, Jens – thanks for being such a wonderfully open and honest person, and for hosting several rounds of pizza roulette, Fabi – roll a D20 on friendship, 20 it’s a critical hit, Pat – either deep conversations or senseless messages, no in between and that’s how it ought to be, Adam – every time we meet, you’re in a good mood, I

aspire to that, Jan – although we see each other rarely, when we do, it feels like you still live next door, thanks for this wonderful lifelong friendship. Gerrit and Tobi, can an evening be spent better than catching up with you two and play a couple board games? Well, maybe on a round trip through Italy with both of you, thanks for being there. The best strategy for recharging after intense times of work has always been visiting home, spending time with my family, my nephews and disconnecting from the world of research for a short time. I can't put my gratitude into words, Mam and Pap, but I'll try. You made it possible for me to go out and find my own way, although this was at times difficult for me, you support me in all that I do and stand strong when the waters get rough, thank you for always welcoming me with open arms and hearts and fostering my curiosity since I was a little child. Here's to you Doris and Joachim. Bruderherz Dani, and also you Carina, every time I spend time with you, it reminds me of the importance of family, being accepted, valued and being important, not the least through Moritz, Matti and Marius who continually remind me of the wonders of the world still to be explored. Reiner, Andy, although I didn't become a dentist, maybe this still counts? Michelle, finding you was the greatest gift of all, I did not know life could be this colourful. Thank you for standing at my side, for being my pillar to lean on when I'm weak, a place of calm when life is too much and bringing light into all of my days.
Juli, you shaped the person I am today, I miss you and hope you'd be proud of me.

Alles geben.

Für meine Brüder Julian und Daniel

Contents

List of Tables

List of Figures

Part I

Background

1 Introduction and motivation

This thesis is centered around the necessity of best performing semantic segmentation algorithms requiring a lot of costly pixel-wise annotations which are hard to acquire, specifically in domains where highly skilled experts annotate. By design prior annotation-cost efficient learning paradigms restrict the expert annotators to specific annotation types, be it few pixel-wise annotations combined with unlabeled data or fast to generate weak annotations, either way forcing them into an inflexible process. In order to enable semantic segmentation technology to be deployed in expert-driven domains, this thesis, presents algorithms that put annotators into the center by dropping the restriction of a single specific annotation type. By accepting broad combinations of imaging data with strong-, weak- and no semantic annotations, our semi-weakly supervised algorithms increase the flexibility of experts in the annotation phase to better fit their needs and the domain at hand. Our developed algorithms offer key solutions for adapting the success formula of semantic segmentation to the small-dataset, expert-driven long tail of computer vision applications.

The economic impact unlocked by artificial intelligence (AI) is estimated *between $3.5 trillion and $5.8 trillion in value annually across nine business functions in 19 industries* [1] while science news postulates *AI is changing how we do science* as it might *spot new particles, see galaxies, sense the public mood from social media* [2]. Arguably one of the biggest disconnects between this euphoric view of AI as supported by recent breakthroughs in AI research [3, 4, 5] can be summarized by a small insight from AI expert Andrew Ng:

I once built a face recognition system using about 350 million images. But when I asked people in the manufacturing industry how many images they had of each defect they wanted to recognize, 50 or fewer was the most common answer. [6]

To put it simply: the long tail of applications for AI in industrial-, scientific- and also medical contexts do not have the data richness that enabled prior breakthroughs, yet they bare immense value to be salvaged which is needed to live up to the euphoric vision of AI.

In this thesis, we consider a portion of this long tail, specifically we consider applications that are enabled by precisely delineating semantic structures in images. The AI or, to get more scientific, the machine learning techniques that are able to fulfill this task are called semantic segmentation algorithms. By solving the task of semantic segmentation it becomes possible to automatically measure shapes, sizes, boundaries and occurrences of all entities within an image as well as to quantify distances between-, contact points among- or overlap of semantically different entities. As might be evident by this variety of possibilities, automatic semantic segmentation is useful for applications from health care, *e.g.* quantifying the progression of a tumor in magnetic resonance imaging [7] or measuring the accumulation of fluid in the retinal layers via optical coherence tomography [8] over industrial- to scientific applications, *e.g.* counting the contact points different cell organelles have to each other in electron microscopy images to gather novel scientific insights [9]. When shifting the focus from applications in the natural image domain of objects, street-scenes or persons where the most notable computer vision advancements are made [5, 10, 11, 12, 13], towards the long tail of applications, the diversity of imaging modalities increases towards other sensors and thereby other imaging properties. With a larger and larger visual difference to common natural imagery, which we all encounter on a day-to-day basis, the imaging modalities on the long tail are not as easy to interpret anymore. Here, highly skilled experts in the respective fields are needed to provide semantic annotations [32] as it is necessary to be familiar with very nuanced structures and work with volumetric-, hyper-spectral- or noisy imaging modalities which often requires a lot of experience or extensive scientific training.

As the pool of possible annotators shrinks to a set of busy experts which have only limited time to apply their expertise to annotate images, naturally, datasets on this

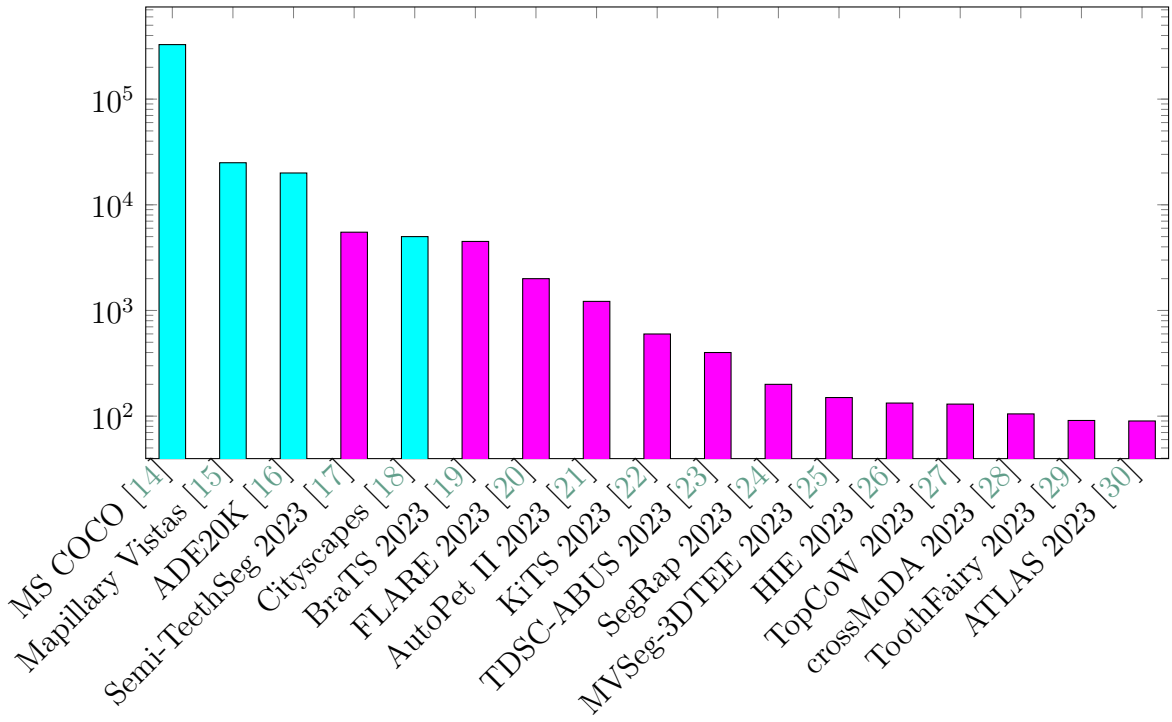


Figure 1: Number of available annotated samples in medical datasets is much smaller due to data properties (*e.g.* volumetric) and due to the necessity for experts needed in annotation. Well-known natural image segmentation datasets (cyan) and most recent medical datasets (magenta) registered as MICCAI 2023 segmentation challenges [31].

long tail are comprised of far fewer annotations as compared to the natural image domain, where scaling annotation through crowd sourcing the annotation process is possible [14]. In Figure 1, we display the number of samples in well-known segmentation datasets from the natural image domain and the most recent segmentation datasets from medical challenges on a log scale. It is evident, that a lot of potential applications in this domain where doctors need to be involved in the annotation process or, the data includes volume data, are heavily annotation constrained.

Domains where process experts, skilled workers or trained scientists are critical in the annotation process will subsequently be referred to as *expert-driven domains*. In

the following chapters, it will be explored how the intricacies of expert-driven domains can be accommodated better in training semantic segmentation algorithms, namely, how we can work with scarce annotation resources more efficiently and how the expert's time should be spent to get further with less of their time. Our goal is to offer an alternative pathway to train segmentation algorithms for use-cases where thousands of costly, pixel-wise annotated images are simply out of reach. Therefore, we want to pivot from the restrictive, time-consuming and annotation-centric supervised semantic segmentation paradigm towards, what we will show to be annotation-efficient and expert-centric, *semi-weakly supervised semantic segmentation*.

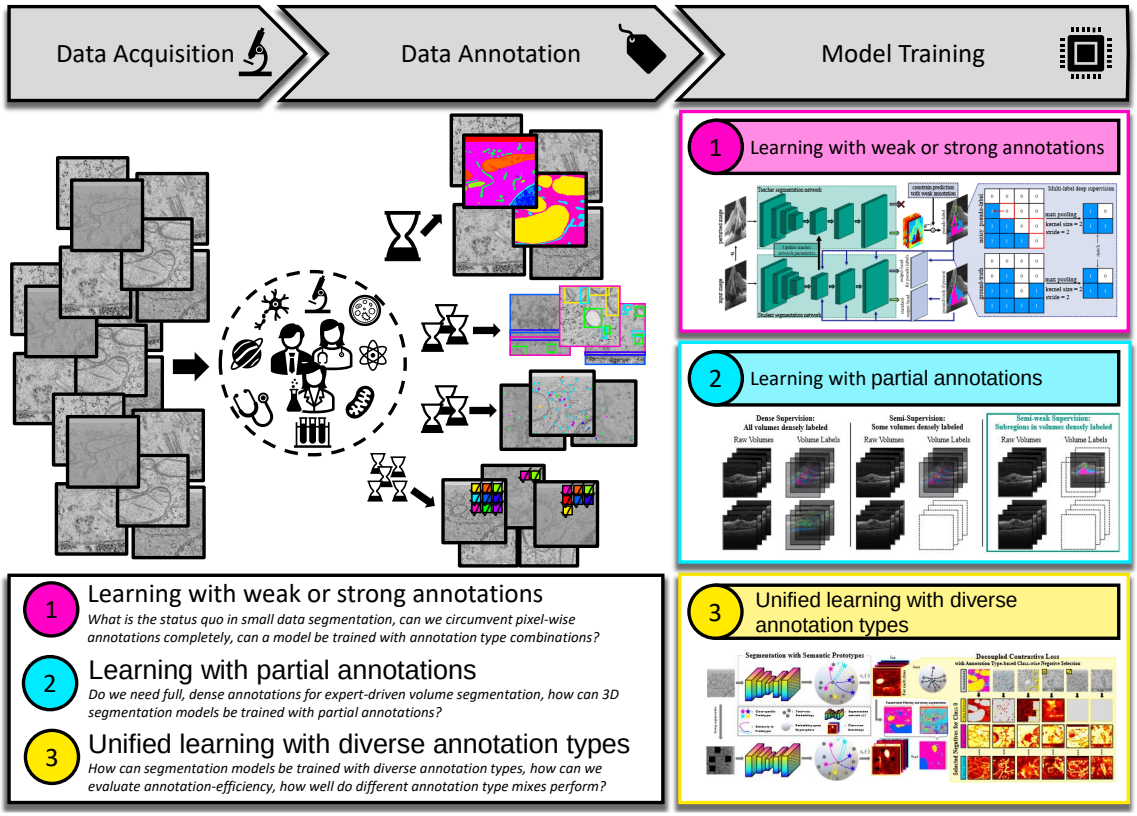


Figure 2: Overview of an expert-centric semantic segmentation pipeline where experts are not restricted to a single annotation type. We display the annotation-efficient and expert-centric segmentation contributions in this thesis on the right.

2 Contributions in this thesis

As described, the standard procedures for training semantic segmentation models were set up mainly with the natural imaging domain in mind, where they have been wildly successful. In this thesis, we explore, how to alter the training procedure of segmentation models for expert-driven domains in order to better respect their critical resource which is the expert's availability to annotate. The contributions are structured into three parts, where segmentation models are trained with (1) pairs of weak and strong annotations, (2) partial annotations and finally with (3) diverse strong-, weak- and partial annotations simultaneously, in order to free the experts from having to provide only a fixed annotation type towards enabling them to provide whatever they have the time for. All these investigations consider scenarios where extremely few, *i.e.* often just a handful of images are associated with pixel-wise annotations, and the size of the datasets fall far behind those in the natural imaging domain. A general overview of the contributions in this thesis is displayed in Figure 2. Implementation of all subsequent contributions was done by Simon Reiß.

2.1 Learning with weak or strong annotations

A first step towards better segmentation in expert-driven domains is presented in chapter 4, which is based on a *CVPR 2021* publication [33]. There, we explore the following research questions:

How far can current segmentation models go with very few pixel-wise annotations in expert-driven domains?

In order to gather insight into segmentation model training in expert-driven domains and to get an idea of the model's behaviour under excessively small amounts of pixel-wise annotations, we first assess how current segmentation algorithms perform in these conditions. We carry this investigation out on the medical imaging domain of optical coherence tomography where ophthalmologists are the experts

needed for annotating different types of retinal fluids within a patient's retina. By rigorously carrying out model training on multiple cross-validation splits for more stable results, and exploring scenarios with varying amounts of annotations from merely one to eight pixel-wise annotations per retinal fluid class, we uncover how the performance of segmentation algorithms progresses with the addition of more and more annotations in generally annotation scarce scenarios. With this insight, we ask:

Are weak annotations sufficient, or are pixel-wise annotations a vital kick-start for training segmentation models in expert-driven domains?

This next question revolves around the prospect to circumvent the high costs of pixel-wise annotations completely by instead working exclusively with more cost efficient weak annotations, such as bounding boxes in expert-driven semantic segmentation. Here, we'll discuss and gather insight into properties of completely weakly supervised approaches in expert-driven domains and see how a naive weakly supervised approach compares to models that have access to very few costly pixel-wise annotations. Naturally, a new question poses itself: Can we get the best of both worlds?

How can we leverage pairs of different annotation types in training a segmentation model and does it help?

Weak labels such as image-level descriptions are fast to generate, costly pixel-wise annotations on the other hand give meaningful cues regarding the delineation of semantic regions. In a first attempt, we look into the prospect of integrating pairs of annotation types to boost the segmentation performance while not having to rely only on costly annotations. This exploration is the starting point towards a more expert-centric training design, adding the possibility for domain experts to annotate on different granularity levels, respecting the expert's time constraints.

2.2 Learning with partial annotations

Expert-driven domains often encompass very different imaging modalities, as compared to the natural image domain and these modalities come with specific properties. It is quite common to be faced with imaging data, that is captured not only in a two- but in a three dimensional manner [7, 8, 19]. Evidently, for each imaged sample the pixel-wise annotation cost scales with the extra dimension which of course adds to the predicament of expert-driven domains. In chapter 5, which is based on an *ECCV 2022* publication [34], we explore how to address this issue:

For expert-driven volume segmentation, do we need full densely annotated volumes to train segmentation models?

Depending on the expert-driven domain that is analyzed, it might only be possible to fully annotate merely a handful of volumes pixel-wise, or due to the excessive size of each volume it might not be feasible to annotate a whole volume at all. Therefore, we try to give insight into the question, whether to train 3D segmentation models, we actually need densely annotated volumes. To investigate this, we analyze how well current models are suited to work with partial annotations where only small regions of the whole volume are annotated. We do this for the medical expert-driven domains of retinal fluid segmentation in optical coherence images and brain tumor segmentation in magnetic resonance imaging. Here, we shed light on extremely scarce annotation scenarios with fewer than ten partial, pixel-wise annotations per class.

How can volumetric segmentation models be trained more effectively using only partially annotated volumes?

With the insight from the prior research question we identify the need for improving segmentation results in such volumetric segmentation scenarios and to make cost-efficient partial annotations a good fit for their training. Therefore, we explore ways to better utilize them while also leveraging the remaining unlabeled regions in each volume for training. By enabling training with cheaper partial annotations, we end

up with a cost-effective 3D segmentation solution, that supplies further flexibility to expert annotators as they can more flexibly use their time on annotating diverse regions, covering more volumes instead of being restricted to annotating a small subset of volumes fully, which often includes annotating redundant adjacent regions.

2.3 Unified learning with diverse annotation types

In chapter 6, which is based on a *CVPR 2023* publication [35], we consolidate the insights into learning with different annotation types that we gatherered in the prior chapters. Here, we put forward a solution to combine a wider variety of annotation types into a single training strategy yielding an annotation-efficient and expert-centric solution for training segmentation models in expert-driven domains:

How can segmentation models be trained with diverse annotation types?

We naively explored the combination of pairs of annotation types in chapter 4. For a truly expert-centric training setup, we would like to be even more flexible and leave the choice of annotation type, be it partial, location-cue free or dense, to the experts and the time they have to annotate. To achieve this, we put forward a strategy to train segmentation models with four diverse annotation types as well as unlabeled images – unifying learning from semantic signals, thereby making possible an expert-centric segmentation pipeline. We evaluate this paradigm in the expert-driven domain of cell organelle segmentation in focused ion beam imaging.

How can semi-weakly supervised semantic segmentation algorithms be analyzed more systematically regarding their annotation-efficiency?

With the higher diversity in annotation types that the models are trained with, the need for a more systematic evaluation procedure of semi-weakly supervised segmentation algorithms arises. To achieve this, we propose a combination of rigorous cross-validation paired with an exponential reduction in expensive pixel-wise annotations.

With this setup, it is easier to gather insight into the vital point at which adding more pixel-wise annotations comes only with diminishing returns in performance. By successively substituting pixel-wise annotations with cheaper annotations, we can measure how different annotation types influence the segmentation performance.

How well do different annotation type mixes perform?

With a semantic segmentation algorithm that is designed to work with arbitrary mixes of annotation types in training and with the tools to better measure annotation-efficiency, we are able to explore the effects of using different annotation type mixes with respect towards the performance. With this we make the first steps towards showing what annotation type mixes give an advantageous cost-performance trade-off which before this work was not possible due to a lack of such a segmentation training strategy. With our novel insights we impact the way that segmentation datasets can be setup and the way segmentation models can be trained enabling practitioners to rethink their procedures from an expert-centric perspective.

3 Related work

When trying to solve the task of semantic segmentation with few pixel-wise annotations, a lot of literature comes prior to and inspired this work. Here, we outline the prevalent paradigms in the research field and how they relate to contributions in this thesis. Specifically, we show what has been done in semi-supervised- and weakly supervised learning which are the most prominent paradigms for considering how to reduce the labour in the annotation process. We also summarize the boundaries to other research directions which are related to ours, but can be seen either as orthogonal or put emphasis on other aspects in their motivation and goal.

In the standard setting, semantic segmentation is approached from a fully supervised

perspective, where it is assumed that for each image in a training set, its dense, pixel-wise annotation is given [36, 37]. In order to more effectively train neural network-based models in this setting, there have been increasingly large efforts to annotate bigger and bigger datasets with dense annotations [18, 14, 16, 38], which has fostered progress in finding new, better segmentation network architectures [39, 40, 41, 42, 43, 44, 45]. While early architectures such as the fully-convolutional network [43] closely followed its classification convolutional neural network (CNN) counterparts [11, 46, 47], subsequent architectures also ensured to be compatible to prominent classification architectures [48, 49] but considered the task of segmentation from an architecture design standpoint more closely. One idea employed by the successful DeepLab [40, 41] architecture was to alter the convolution operation in order for it to cover larger receptive fields via dilated convolutions [50]. Enabling the networks to capture long ranging contexts in images, attention mechanisms [51, 52] were frequently introduced into segmentation architectures [53, 54, 55, 56], leading up to self-attention [57] based segmentation networks [58] which often build upon visual transformers [49] or swin transformers [59] as backbone. This progress in architecture development boils down to several factors, one of which is the availability of a large number of pixel-wise annotated samples [18, 16, 38] as well as strong natural imagery pre-trained architectures [49, 48, 59] and pre-training strategies [60, 61, 62, 63] obtained from training on large datasets [64] which help in common street-scene- or everyday scene-centered segmentation benchmarks. Moving towards expert-driven domains, where experts have to annotate due to the difficulty of interpreting the non-standard imaging data, strategies such as crowd sourcing [14, 64] the annotation process or crawling online sources [65] to scale can not be copied.

As such, in expert-driven domains such as bio-medical imaging, where segmentation datasets are magnitudes smaller [8, 9], the Unet architecture [44] and its variants [66, 67, 68, 69, 70, 71, 72] is still prevalent [73, 74, 75, 76, 71, 77, 9, 78], due to its robust performance with simple training configurations concerning learning rate, scheduling and other hyper-parameters as well as when few annotated images are

available. Although strides towards self-attention-based architectures are also being made in the medical domain recently [79, 80, 81]. Due to its general robustness we often chose the Unet architecture in our experiments, where we need a fast trainable and stable architecture which can handle scenarios where it is only supplied with a handful of pixel-wise annotations and yields strong results when serving as base segmentation network for a wide variety of baseline training strategies. Apart from different visual properties, in expert-driven domains, the imaging data might come in a different shape, such as with more channels or a volumetric shape [82, 83, 8, 9]. To semantically segment volumetric data, methods with 3D convolutional operations where proposed [84, 85] which led up to the adaptation of Unet to 3D Unet [66] which is still among best performing methods in medical challenges [78, 86] and which is also chosen as base architecture for volumetric segmentation tasks in this work. Yet, for volume segmentation, a lot of focus has been to adapt the 3D Unet architecture to different datasets and segmentation tasks by adding multiple pathways [69], self-supervised training regimes [71] or considering boundary regions specifically [72, 69, 70]. The paradigm of training models by utilizing deep supervision has been considered in both standard 2D [87, 88, 89, 90, 91] and in volumetric segmentation [70, 92]. The positive properties in low-data scenarios [70] and on convergence in training as well as generalization and vanishing gradients [93] have led to a lot of methods utilizing deep supervision to inject semantic information into earlier layers of the network [94, 95, 96]. While the idea of deep supervision was introduced for classification [97, 93, 46], especially the expert-driven medical imaging community, plagued by small datasets, took hold of the idea and frequently added it into training schemes [87, 92, 88, 89, 90, 70, 91, 98, 68]. In this thesis, the deep supervision paradigm is made use of within a new semi-supervised method, which we benchmark against some of the semi-supervised algorithms described next.

3.1 Semi-supervised segmentation

Semi-supervised learning, where the training set for machine learning algorithms is made up of a small labeled portion and a, generally much larger, unlabeled portion, is a prominent choice in segmentation scenarios, as small amounts of pixel-wise annotations is commonly what is possible to get hold of. One pathway to integrate both annotated images and unlabeled images into training is by training a network on the labeled portion and inferring so called pseudo-labels for the unlabeled images and continue training with those as well [99, 62, 100, 101]. An approach which bridges the gap between pseudo-label methods and so called consistency regularization approaches is FixMatch [102]. It works with two differently strong augmentations on an image where the weakly augmented image is used to infer the pseudo-label and the strongly augmented image is used as input to train the network with back-propagation and the pseudo-label as ground truth. Originally, this approach was designed for semi-supervised classification but has shown strong performance in segmentation as well [103]. Enforcing similar predictions from an image which was augmented in two different ways is one way of consistency regularization, while others include perturbing the forward pass of a network in different ways, *e.g.* by applying dropout [104], using different network architectures [105] or multiple networks such as a student and teacher [104, 106, 107] or by manipulating the input to the networks [108, 109] using methods like CutMix [110]. In designing our own methods, we take inspiration from student-teacher [104] or siamese setups [111, 112] and add different augmentations in order to force the model to learn augmentation invariance for semantic segmentation, which help us in addressing the problem of data scarcity. Differently augmenting a single image is also used in the realm of contrastive pre-training of classification- [113, 114] but also segmentation methods [115, 116, 117, 118, 119]. The paradigm of contrastive learning is an intriguing one, which we will make use of by disentangling class associations among images with differently granular annotations. Further, we utilize the contrastive paradigm in volume segmentation, to enforce different properties on individual embeddings of input voxels. This, we do in

a setting where we train with partially labeled volumes [120, 66] and unlabeled volumes which is an adaptation of semi-supervised volume segmentation from literature, where it is generally assumed that the labeled portion of volumes consists of densely, voxel-wise annotated volumes [121, 122, 123, 124, 125, 126]. These semi-supervised volume segmentation solutions consider the problem though a variety of lenses, including through adversarial learning [121], through processing multiple views on the 3D data [122], through student-teacher setups combined with uncertainty modeling [123, 124, 125], or via contrastive objectives [127, 128, 129]. Some variants of semi-supervised learning considers graphs [130, 131] in their design, for example to relate different labeled and unlabeled images [132, 133]. In our methodological desgin, at times, we also make use of the view through graphs, yet, we do this in terms of viewing the input volume as a graph, *i.e.* each representation of a voxel as a node in it. This view on the input data as a graph is common in computer vision, *e.g.* in context of post-processing methods such as Conditional Random Fields [134, 135, 136, 137]. While the semi-supervised learning paradigm is a good choice for scenarios where the annotation budget only allows for a small set of annotated samples, it lacks flexibility, as it still restricts the annotators, or in expert-driven domains the busy experts, to sit down and strictly only provide pixel-wise annotations while at some point providing faster, coarse annotations might be already sufficient and might help in covering a higher diversity of samples with semantic annotations.

3.2 Weakly supervised segmentation

Research in making use of coarse annotations to train semantic segmentation models has seen a lot of interest in the natural image domain, there these coarse and fast to obtain annotations include image-level labels [138, 139, 140, 141, 142, 143, 144, 145], scribbles or points [146, 147, 148, 149, 150, 151, 152, 153] and bounding boxes [154, 155, 156, 157, 158]. In early work on training segmentation models from image-level labels [138] the paradigm of multiple instance learning [159] has been made use of, while most of successive work utilized feature attribution methods [160, 161, 162].

For feature attribution methods, a classifier is trained on the image-level labels in a way such that coarse location cues can be extracted for the training images, which then are refined by weakly supervised methods using specifically designed prior assumptions [139, 141, 142, 143, 145]. A few approaches also investigate this paradigm in the expert-driven medical domains, *e.g.* optical coherence tomography [163, 164], yet for the most part the assumptions that lead to good results in the natural imaging domain are not naively transferable to the wildly different data in medicine. In our experiments we investigate the implications of additional image-level labels on a semi-supervised segmentation network.

Bounding boxes which are drawn around the entities to segment can also offer a quick, yet coarse location cue which has been used to train semantic segmentation models with in the natural image domain [158, 157, 165, 155]. Many of these methods address the ill posed problem of weakly supervised segmentation with boxes by integrating algorithms such as GrabCut [166], Multiscale Combinatorial Grouping [167], or Selective Search [168] which can derive a strong initial segmentation from the boxes which can be used for segmentation model training. For image domains where these algorithms lead to good initial segments, these weakly supervised solutions are applicable, yet for expert-driven domains, this might not always be the case due to the different data distribution. Expert-driven medical domains such as positron emission tomography [169] and magnetic resonance imaging [170, 171] have also seen the application of box-based weakly supervised segmentation. Whether bounding boxes can be a substitute for dense pixel-wise annotations in expert-driven domains, or whether a few pixel-wise annotations are disproportionately effective is one question investigated in our experimentation.

The last frequently explored weak annotation types that we discuss here, are partial, incomplete scribble or point annotations [150, 146]. Due to the ease of acquisition, they are a common choice also in the medical field [172, 173, 174, 175, 176]. Angles from which these works approach learning segmentation from scribbles include adversarial objectives using additional unpaired pixel-wise masks [174], bootstrapping

pseudo-labels for histopathology images [175] or working with extreme point clicks for a volumetric segmentation task [176]. Related, to this form of supervision, in our experiments, we train volumetric segmentation models with sparsely annotated volumes and design a method which can better cope with them.
Weakly supervised segmentation has shown strong performance in the natural image domain, what makes it difficult for expert-driven domains where the images have very different properties from natural images is that the designed constraints might not be transferable between these domains. Therefore, specific constraints have to be found again and again for each domain, hindering easy deployment.

3.3 Semi-weakly supervised segmentation

Both semi- as well as weakly supervised segmentation offer valid pathways towards reducing the required amount of densely annotated images as would be needed to obtain good results via supervised learning. A step towards a more flexible learning scenario has been argued for by Choe *et al.* [177], namely training with both weak and dense sets of annotations in a semi-weakly supervised way. This style of training with diverse mixes of available semantic information is sometimes also referred to as mixed- or omni-supervised training [178, 179, 180]. With a segmentation solution in this style of training, it is possible to, from an algorithmic perspective, influence the annotation process, by accepting a varying granularity of annotations at once, in our case, freeing the time-constrained experts to flexibly provide any annotation type that they have the time for – be it a few pixe-wise masks, or additional weak annotations. Early explorations combined pixel-wise masks with image-level information [181] and bounding boxes [165]. Both scenarios using masks with image-level labels [182, 183, 184, 105, 185, 186, 187] and masks with boxes [157, 188, 189, 190, 191] were explored frequently, yet, the former scenario was often investigated only briefly in these works. More seldom annotation mixes explored in literature include the combination of mask annotations with scribbles [146, 192] or unlabeled images with scribbles [178]. The most diverse use of annotation types in segmentation was in Li *et al.*'s work [156] on

bootstraping a panoptic segmentation system, for which at different stages image-level labels, boxes and masks come into play. This thesis aims at quantifying the effect of different semi-weakly supervision scenarios for segmentation to get a deeper insight into which annotation type combinations are most effective in yielding good results at manageable costs for expert-driven domains. This includes methodological contributions we put forward to effectively benefit from diverse semantic annotation types beyond pairs of annotations types as previously done. Thereby we open up possibilities for expert annotators to spend their time flexibly on annotations of different granularity to the biggest effect.

3.4 Orthogonal research fields

This thesis is centered around reducing the effort for annotation in expert-driven domains from an algorithmic, neural network training strategy perspective, where we aim at making the annotation process flexible by accepting a broad variety of semantic cues. Yet, there exist research fields with aligned goals, *i.e.* reducing the annotation effort, that are orthogonal to semi-weakly supervised segmentation and thus could be applied jointly. Next, we briefly mention and outline these fields.

Active learning: In active learning, the main idea is to, before annotation, estimate how much benefit each of the unlabeled examples will have on the model performance when it is supplied with an annotation. For semantic segmentation a variety of approaches exist, from the utilization of adversarial learning [193, 194] to estimation of uncertainty maps [195] which sometimes are computed through multiple augmentations [196]. During the process of active learning, segmentation models are trained, which could also be done in a semi-weakly fashion leveraging our training strategies, which on top offer the possibility to extend the active learning scenario to include suggestions on which annotation type is the most beneficial for each unlabeled image in terms of a performance gain and cost trade-off. To get a better view on results of active learning, intertwining it with semi-supervised methods has even been argued

for [197] and was explored already [198] including in the expert-driven histopathology domain [199], which further validates that methods presented in this thesis are readily applicable.

Interactive segmentation: Rather than decreasing the amount of pixel-wise annotations, which we will investigate in our experiments, in interactive segmentation [200, 201, 202] the goal is to lower the time it takes for each pixel-wise annotation to be drawn. This is done by instead of carefully delineating segments in images, asking the annotator to provide a sequence of clicks, and with each click computing a higher quality mask suggestion. Interactive segmentation is common to the medical domain including for volumetric imagery [203, 204, 205, 206], and can be utilized to lower the time for each pixel-wise annotation before semi-weakly supervised training, or semi-weak algorithms could use the information form sparse clicks as additional supervision signal.

Transfer learning: The idea of transfer learning is to pre-train a model on a large dataset in order to learn general, transferable features which will be useful for fine-tuning the model on a down-steam task where only few annotations are available. Transfer learning has proven useful to reduce the amount of annotations needed in the medical domain [207], where, for classification, the number of annotations could be reduced by 40%, 59% and 70% for mammography [208], chest x-ray [209] and dermatology [207], respectively, without performance degradation. Reductions in annotations for segmentation in expert-driven domains, where the datasets are much smaller than for classification has also been investigated [210, 211, 212] and could easily be combined with the benefits of the annotation-efficient solutions in this thesis.

Annotation budget allocation: This thesis investigates the implications of training segmentation models with different annotation type mixes, which coarsely relates to annotation budget-focused research on investigating whether to annotate images with dense or weak annotations [213] or trying to estimate the data requirements of an algorithm [214, 215, 216]. Combining these works with the methodological

contributions to integrate diverse annotations, which will be presented next, could paint a broader picture and lead to more economical annotation-budget allocation strategies for practitioners.

Part II

Expert-centric Semi-weakly Supervised Semantic Segmentation

4 Learning with weak or strong annotations

Neural network-based semantic segmentation algorithms generally require a lot of pixel-wise annotations to produce strong results. This chapter investigates how effectively these algorithms can be trained when drastically reducing the amount of annotated data and how their performance evolves when successively granting them more and more pixel-wise annotations. Then, current semi-supervised algorithms are explored towards their behavior in this setting, and the paradigm of semi-weakly supervised segmentation is studied as pathway to more flexibly profit from different types of semantic information. Specifically, it is investigated how the performance changes when intertwining pixel-wise annotations with image-level information which provides experts with more options to annotate. Further, naive weakly supervised training is put to the test, in order to investigate whether supplying models with only box information – depriving them from mask information completely – is an economical pathway to kick-start segmentation training, or whether few mask annotations aside unlabeled or weakly labeled images are unreasonably effective and provide the better annotation pathway.

This section is based on a publication in *CVPR 2021* [33].

4.1 Introduction

The availability of massive heaps of labeled training data [64, 14] facilitated by the internet and the parallelism from specially designed computing hardware in combination enabled more extensive and efficient training of deep neural networks, bringing large leaps in performance with them [11, 47, 48]. This led neural networks and their training via error back-propagation [217] to be front-runners on a diverse variety of computer vision tasks, from classification [11, 47, 48, 46], to detection [218, 219, 220] and semantic segmentation [43, 44, 40, 41] to name some of the most prominent. To

consolidate this progress into other domains than internet-driven imagery, for example into medical imaging, one major hurdle presents itself: the effort of setting up new and large enough annotated datasets for the training data requirements of deep neural networks. For setting up a semantic segmentation dataset in *e.g.* the medical domain not only are time-intensive, pixel-wise annotations covering the whole image needed, but medical experts need to provide them as it requires a lot of experience and training to grasp the nuances of pathology and abnormality in medical scans [32]. Yet, these highly skilled experts do not go through the long training to develop their abilities for annotating images with precise semantic location information, but, in case of medical experts, to spend their time on quite literally saving lives. This results in a small time-budget based on the expert's availability for providing annotations, putting heavy restrictions on the number of annotated samples the semantic segmentation algorithms are trained with, often far below the size of natural image datasets with multiple thousand pixel-wise annotations.

The following chapter is an investigation into the performance of semantic segmentation algorithms when they are supplied only with a handful of pixel-wise annotations in the expert-driven medical domain of retinal fluid segmentation in optical coherence tomography scans [221]. Specifically, a variety of semi-supervised semantic segmentation algorithms are trained with extremely few pixel-wise annotations and some unlabeled retinal scans. In order to better grasp how big the requirement for pixel-wise labels for acceptable segmentation performance actually is, the amount of pixel-wise annotations is increased successively in an effort to close the performance gap as compared to training with the full pixel-wise annotated dataset. Thereafter, we investigate how well a naive weakly supervised approach which uses bounding boxes as most time-consuming annotation performs, and whether skipping costly pixel-wise annotations completely in favor of boxes is a valid option. To take a first step towards uncovering the implications of learning from annotation type mixes, all experiments are repeated using not pixel-wise and unlabeled retinal scans but pixel-wise and image-level labeled scans, probing into the semi-weakly supervised

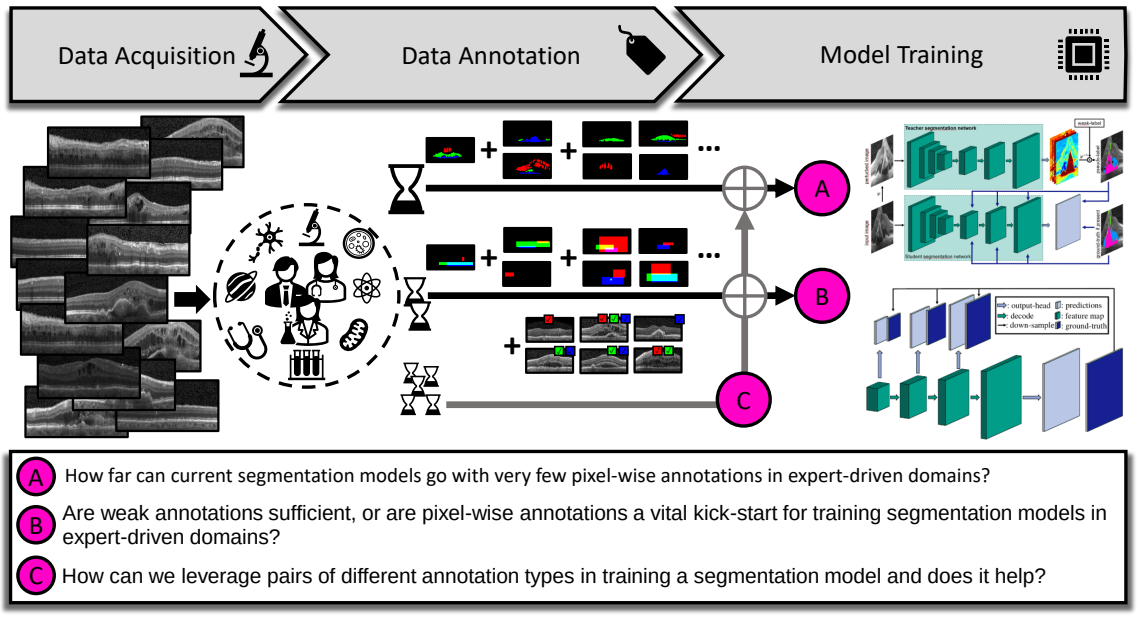


Figure 3: Overview of the main research questions in this chapter, they will be explored on an optical coherence tomography dataset [8] where medical doctors are needed in the annotation process. Further, the *Mean-taught Deep Supervision* method is outlined which can help in annotation-scarce, expert-driven domains.

segmentation paradigm and quantifying its segmentation performance. All these experimental settings are carried out with a rigorous ten-fold cross validation evaluation protocol to reduce the performance fluctuations that are expected in scenarios with extremely few annotations.

Ahead of these investigations, the scenario of semi-weakly supervised retinal fluid segmentation is formally introduced in Section 4.1.1 and the algorithmic solution to this task, the *Mean-taught Deep Supervision* model, which can be trained in a fully supervised, semi-supervised and semi-weakly supervised manner is presented. It is motivated by insights from semi-supervised learning [104], extended with the idea of deep supervision and a first pathway in this thesis for integrating different annotation types at once. The three main research questions of this chapter are summarized visually in Figure 3.

4.1.1 Problem statement

Next, the task of semantic segmentation is outlined, which we aim at tackling through semi-weakly supervised learning. Generally, semantic segmentation algorithms are trained using an image dataset:

$$\mathcal{D} = \{x_1, ..., x_n | x_i \in \mathbb{R}^{3\times H\times W}\} \ . \tag{II.1}$$

For classical supervised learning, each image x_i in dataset $\mathcal{D}$ is associated with a dense annotation, or mask m_i:

$$\mathcal{M} = \{m_1, ..., m_k | m_i \in [0,1]^{(C+1)\times H\times W}\} \ . \tag{II.2}$$

Note, for $m_i \in \mathcal{M}$, we assume that at each spatial position $(x, y) \in H \times W$ only a single class is set to one (one-hot encoding). A segmentation algorithm makes use of these rich resources $\mathcal{D}$ and $\mathcal{M}$ to train a model in order to, for a novel image x_t that has not been seen before in training, compute a segmentation $s_i \in [0,1]^{(C+1)\times H\times W}$ which correctly categorizes each pixel into one of the $C + 1$ classes.

As we described earlier, in expert-driven domains, acquiring a dataset with pixel-wise masks $\mathcal{M}$ is very labour-intensive and heavily dependent on the experts availability to annotate. The paradigm of semi-supervised semantic segmentation can be used even when only a subset of the images in $\mathcal{D}$ are associated with masks, *i.e.* when $|\mathcal{M}| < |\mathcal{D}|$. In semi-supervised learning a usual assumption is that the unlabeled portion is much smaller than the labeled portion ($|\mathcal{M}| \ll |\mathcal{D}|$).

The second paradigm that addresses the issue of dealing with the limited time of annotators is the paradigm of weakly supervised learning. Here, instead of integrating unlabeled images or learning from costly pixel-wise annotations, segmentation algorithms are trained on a dataset $\mathcal{D}$ which is associated not with masks but with

weaker and thereby easier to acquire annotations:

$$\mathcal{B} = \{b_1, ..., b_n | b_i \in [0,1]^{(C+1)\times H\times W}\} \ , \tag{II.3}$$

$$\mathcal{I} = \{l_1, ..., l_n | l_i \in [0,1]^{(C+1)}\} \ . \tag{II.4}$$

Common choices for weaker annotations are bounding boxes $\mathcal{B}$, which give coarse location cues covering the whole extent of an entity or completely location-less cues such as global image-level labels $\mathcal{I}$ that merely indicate the presence or absence of a class. We replace the common representation of bounding boxes as two points $(x_1, y_1), (x_2, y_2) \in H \times W$ by a mask-like notation, where a foreground class c is set to one at all spatial positions that fall inside one of the bounding boxes of class c. As a special case, the background class $C + 1$ is set to one at all left empty regions. Even though the training annotations change for weakly supervised segmentation, the goal still is to infer a correct pixel-wise prediction for unseen images.

Both semi- and weakly supervised segmentation offer valuable options towards modeling learning and towards designing annotation pipelines. What we are interested in is profiting from both directions as we (1) want to profit from pixel-wise annotations and potentially unlabeled images but (2) also gain the option to additionally learn from weak annotations. Therefore, we define the semi-weakly supervised segmentation scenario as training a segmentation algorithm based on the dataset $\mathcal{D}$ and associated annotations consisting of any subset of $\mathcal{M}$, $\mathcal{B}$, $\mathcal{I}$, where each of these annotation type sets may only cover a portion of $\mathcal{D}$. In experiments of this chapter, we make a first assessment of semi-weakly training and investigate the effects of learning from annotation type combinations: $\mathcal{M} + \mathcal{I}$ and $\mathcal{B} + \mathcal{I}$. There, images which are not supplied with a mask m_i or a box b_i are annotated with an image-level label l_i.

4.1.2 Preliminaries

Next, some general notation is introduced in order to, more concisely, write about methodological intricacies in the following sections.

Image processing First, we outline how an image is processed via an encoder-decoder segmentation architecture. Throughout the segmentation network, there exist feature maps $f \in \mathbb{R}^{d \times H \times W}$ after all convolutional, normalization and activation function computations when processing an image. We refer to the leading dimension as the number of feature channels, the following two dimensions H and W to the spatial dimensions or the spatial extent of feature maps. Encoder-decoder segmentation architectures generally process an image via an encoder part which successively spatially compresses the image from the dimensions $3 \times H \times W$ into a feature map $f_0 \in H_0 \times W_0$. The decoder structure in encoder-decoder architectures then successively up-scales the feature map f_0. This repeated up-scaling leads to a sequence of intermediate feature maps $f_0, \ldots, f_h$, with the outermost feature map $f_h \in \mathbb{R}^{d \times H \times W}$ and with the property that for a feature map $f_i \in \mathbb{R}^{d \times H_i \times W_i}$ we assume that:

$$\forall_{i \in \{0,\ldots,h\}; i<j} : H_i \leq H_j \ \wedge \ W_i \leq W_j \ . \tag{II.5}$$

Thus, the feature maps in the decoder monotonically increase in the spatial dimensions. In our experiments, the feature maps' spatial sizes decrease by a factor of two after each encoder-block and repeatedly double in size in the decoder. The feature maps in the decoder of the segmentation model will later be utilized to integrate a deep supervision learning signal. To end up with a pixel-wise prediction the segmentation network produces a feature map f which is transformed by an output-head containing $C+1$ 1×1 convolutions to map it onto the number of classes to segment. We adapt this slightly and formulate output-heads $\kappa(\cdot)$ as a sequence of 1×1 convolution, batch normalization [222] and ReLU non-linearity [223] followed by a final 1×1 convolution. Computing predictions based on a given feature map f_i will be referred to as $\kappa_i(f_i) \in \mathbb{R}^{C+1 \times H_i \times W_i}$, with $\kappa_i(\cdot)$ denoting an output-head for a specific feature map resolution in the decoder. Further, some of our models work with multiple output-heads at the outermost layer, which leads to an added output-head next to the standard output-head $\kappa(\cdot)$. This setup is commonly employed in multi-task architectures [224, 225].

Supervision signals To optimize the encoder-decoder network towards a set of parameters that gets closer to solving the semantic segmentation task, loss functions are needed in order to evaluate the discrepancy between the current output of the model with respect to the correct segmentation. Steering the parameters into the correct direction via stochastic gradient descent for images where mask annotations are present can be done by minimizing the pixel-wise cross-entropy loss:

$$\mathcal{L}_{CE}(\kappa, f, m) = -\frac{1}{H \cdot W} \sum_{i,j,c=1}^{H,W,C} m^{c,i,j} \cdot \log(\alpha(\kappa(f))^{c,i,j}) \,, \tag{II.6}$$

here, $\alpha(\cdot)$ is the softmax function applied along the first dimension.
As hinted at earlier, we aim at introducing intermediate supervision signals via deep supervision into the network training. For this, we will need a cross-entropy loss formulation which allows for the presence of multiple classes at the same spatial location, which leads us to the commonly used binary cross-entropy loss:

$$BCE(o, t) = t \cdot \log(\sigma(o)) + (1 - t) \log(1 - \sigma(o)) \tag{II.7}$$

$$\mathcal{L}_{BCE}(\kappa, f, m) = -\frac{1}{H \cdot W \cdot C} \sum_{i,j,c=1}^{H,W,C} BCE(\kappa(f)^{c,i,j}, m^{c,i,j}) \tag{II.8}$$

This loss most commonly uses a sigmoid normalization $\sigma(\cdot)$ which we will also make use of in the ensuing section.

4.2 Multi-label deeply supervised networks

In the following sections, we show a specific technique to train networks in a flexible semi-weakly supervised scheme which we term *Self-taught Deep Supervision* and extend it to the so called *Mean-taught Deep Supervision* training setup. These techniques will enable the investigation of the research questions of this chapter, as they can be trained with a varying annotation type mix. But first, we introduce a simple trick to add a semantically consistent deep supervision variant into the network

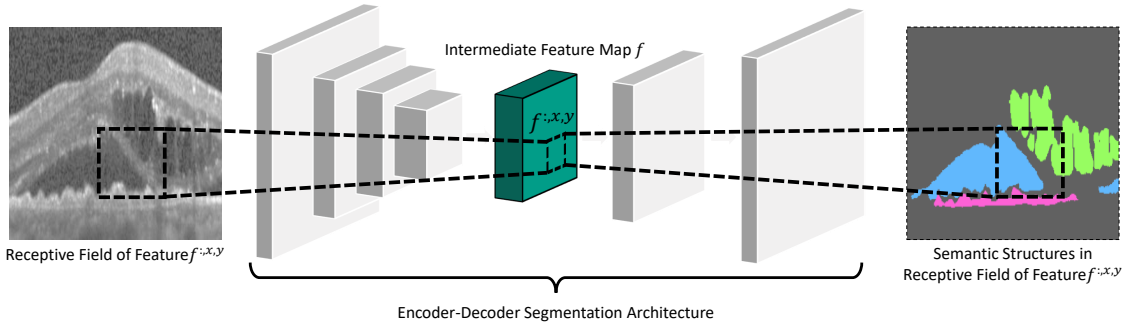


Figure 4: Intermediate features can encompass a multitude of semantic classes within their receptive field and thus can be considered descriptors of these image patches. This insight is the founding consideration for the *Multi-label Deep Supervision* loss.

training, the *Multi-label Deep Supervision* loss formulation, which both semi-weakly supervised training schemes make use of.

4.2.1 Multi-label deep supervision

The idea of deep supervision was first introduced for image classification [93] with the problem formulation to merely identify a single class in each image. To integrate a learning signal, the intermediate feature maps were pooled spatially and associated with a layer-specific classifier, the predictions of which can be used to calculate a loss. In semantic segmentation, where the locality of semantic structures is front and center, retaining the spatial extent of the feature maps enables a sensitivity towards location through the deep supervision learning signal. As the intermediate feature maps, which are supplied with a deep supervision learning signal, are smaller $(f_0, \ldots, f_{h-1})$ than the full-scale ground-truth this spatial mismatch has to be addressed. Apart from [226] which use a nearest-neighbor interpolation to down-scale the ground-truth annotation and thereby losing semantic information, most commonly the feature map is up-scaled [88, 89, 90, 92] using an interpolation strategy or learned up-scaling and subsequently adds an output-head on top of the up-scaled feature map to end up at full-scale pixel-wise predictions.

This second way of up-scaling the intermediate feature map and segmenting it is

a quite hard task, specifically, it is the same task that the entirety of the network tries to achieve. Therefore, designing deep supervision in such a way forces the network to, for an exemplary intermediate feature map $f_{small} \in \mathbb{R}^{d \times 10 \times 10}$ and a corresponding ground-truth mask $m_{big} \in \mathbb{R}^{c \times 100 \times 100}$ to up-scale each feature $f_{small}^{:,x,y}$ at the spatial location (x, y) to infer a complete patch of size 10×10 to accommodate the big ground-truth annotation. This way of modeling has the shortcoming that the network itself hast to learn an up-scaling at each intermediate feature map and thereby the intermediate features need to convey complex information about the fine-grained spatial relations in the full-scale output space which in the small feature maps after the encoder of the encoder-decoder architecture might be hard to unravel.

To circumvent these challenges, we take the route of [226] and down-scale the ground-truth. Yet, we want to achieve this without loosing semantic information in the process, which the nearest-neighbor interpolation choice of down-scaling does, as it only allows for one class at each spatial location. In Figure 4, we visualize the motivation for our approach. There, for a feature vector $f^{:x,y}$ at a given spatial position (x, y) in an intermediate feature map f we can display its receptive field, *i.e.* highlight which pixels from the input image went into the computation of $f^{:x,y}$. When overlaying the receptive field of the feature $f^{:,x,y}$ with the corresponding ground-truth mask (Figure 4 right hand side), we see, that $f^{:,x,y}$ should achieve to encode all semantic classes contained in the overlaid field. With this view, individual features in intermediate layers can be seen as patch descriptors with regard to the input image. To make sure that these individual features capture the semantics contained in the patch that they describe, we propose to enforce a multi-label loss with the target for a feature vector $f^{:x,y}$ consisting of all present semantic classes in its receptive field. As we aim to do this by down-scaling the ground-truth, the model does not have to learn an up-scaling procedure in the intermediate output branches which also leads to a reduced learnable parameter count.

With the formulation of ground-truth masks as binary tensors (II.2), producing the

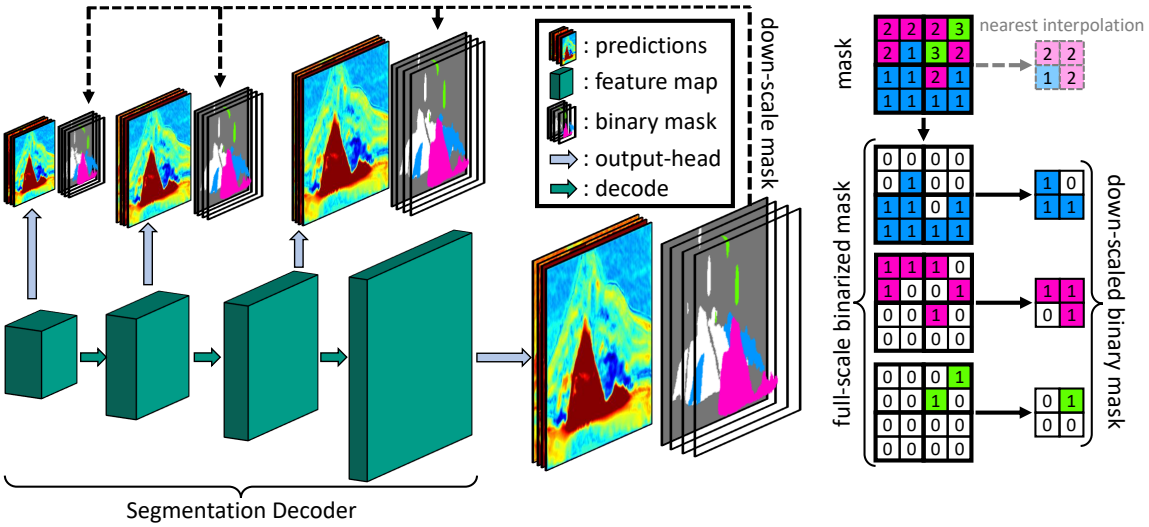


Figure 5: The process of integrating segmentation mask information into intermediate layers in the decoder of a segmentation architecture utilizing our *Multi-label Deep Supervision* mechanism. Using max-pooling on multi-label binary masks retains semantic information as displayed to the right.

down-scaled multi-label ground-truth can be achieved efficiently by applying a max-pooling kernel on top of the binary ground-truth with a fitting kernel-size and stride to match the individual feature map's spatial extent. This is a semantic preserving way of down-scaling the mask annotation, as the occurrence of a class is not lost in the interpolation procedure. We refer to the down-scaled target for an intermediate feature map f_i as $m_i^* \in \mathbb{R}^{C+1 \times H_i \times W_i}$, where the multi-label binary target for the patch descriptor $f_i^{:,x,y}$ is simply $m_i^{*\ :,x,y}$ as it encompasses the aggregated semantic classes of the associated patch from the input image. To integrate the down-scaled ground-truths, we use separate output-heads $\kappa_i(\cdot)$ which convolve over the intermediate feature map f_i and produce a prediction $\kappa_i(f_i) \in \mathbb{R}^{C+1 \times W_i \times H_i}$ of the same spatial extent as f_i. These output-heads are applied throughout the decoder part of the segmentation architecture as displayed in Figure 5, and are supplied with the semantic preserving down-scaled targets as shown on right hand side. To train these intermediate segmentation output-heads, the following multi-label segmentation loss

function is applied:

$$\mathcal{L}(f_1, \ldots, f_h, m_1^*, \ldots, m_h^*) = \frac{1}{h} \sum_{k=1}^{h} \mathcal{L}_{BCE}(\kappa_k, f_k, m_k^*). \tag{II.9}$$

We term this loss function *Multi-label Deep Supervision*, as it reformulates the deep supervision integration into binary class-wise predictions to preserve all semantic classes within the receptive field at each spatial location in the hierarchical feature maps of the segmentation decoder. Our training with this loss is paired with a standard cross-entropy loss function as in Equation (II.6) on the outer-most layer which also serves as prediction output-head for segmentation inference.

4.2.2 Self-taught deep supervision

Integrating Multi-label Deep Supervision into the network training naively only extends supervised learning with an additional loss term. Yet, what we can make use of to leverage it in a semi-supervised fashion is it's combination with pseudo-labels [99]. This is motivated by the idea that inferring noisy labels for images without associated masks via the segmentation network itself offers a way to make use of unlabeled images. Self-inferring the labels for images will often lead to a faulty segmentation which then is still used to train with. Yet, if we integrate the noisy Pseudo-labels using the Multi-label Deep Supervision formulation, by down-scaling them, small inaccuracies can be smoothed out and features, *i.e.* patch-descriptors within intermediate layers leverage the smoothed pseudo-label version which can better match the unavailable ground-truth, as the toy example on the right side of Figure 6 showcases. This loss design directly enables learning the semantic segmentation from noisy pseudo-labels at very coarse granularity (small scale) and successively a more refined segmentation in later decoder layers which have to capture more detailed semantic structures up to the full resolution.

With these considerations, we introduce the *Self-taught Deep Supervision* training strategy and architecture, which is trained on masks in a supervised fashion with the

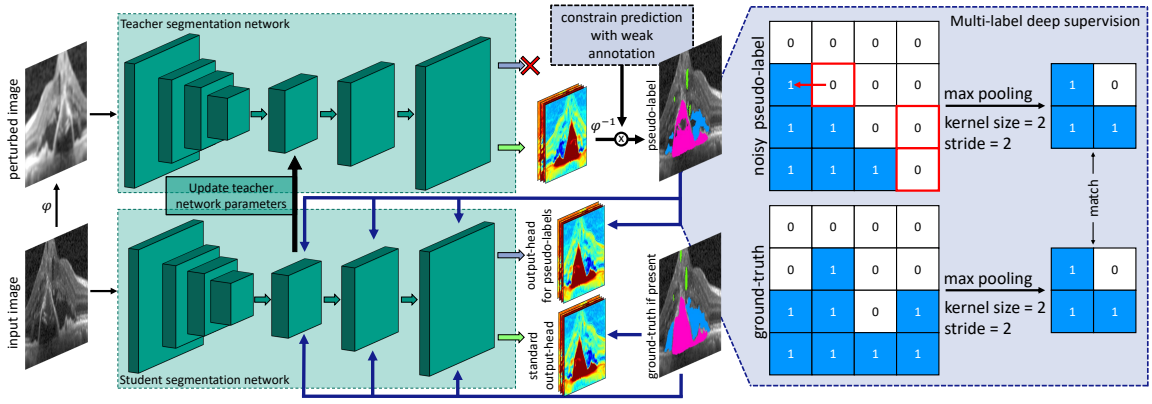


Figure 6: The proposed approach combines pseudo-labeling via a mean-teacher with a novel perspective on deep supervision. By perturbing the input to the teacher and reversing geometric transformations in output-space, we streamline mean-teachers for segmentation. The new deep supervision, *i.e. Multi-label Deep Supervision* introduces a smoothing effect for noisy pseudo-labels: At smaller scales, erroneous predictions (red) like small shifts or few missed pixel-classifications get smoothed out.

cross-entropy and Multi-label Deep Supervision losses. To enable semi-supervised learning we infer a pseudo-label $\hat{p}$ for an unlabeled image from it's corresponding outermost feature map f_h online using the network prediction $\kappa(f_h)$ itself:

$$\hat{p} = \sum_{i,j}^{H,W} \arg\max_{c} \alpha(\kappa(f_h)^{c,i,j}) \quad , \tag{II.10}$$

which assigns the class with the highest classification score to a pixel. Afterwards, we form binary tensors from the pseudo-label $\hat{p}$ and insert it into the Multi-label Deep Supervision loss. On the outermost, full scale feature map f_h we utilize the standard output-head $\kappa(\cdot)$ for the clean, accurately annotated images coupled with the cross-entropy loss and a second output-head $\kappa_h(\cdot)$ for the pseudo-labeled images as part of the Multi-label Deep Supervison loss. This design has experimentally shown to work better, specifically, when the clean output-head is utilized for computing pseudo-labels as well as in inference. Dual-head architecture designs with stop-gradients (ours: implicit stop-gradient via pseudo-labels) have also been found to be crucial

for producing non-collapsing representations in self-supervised literature [111].
When moving from the semi-supervised training scenario to training in a semi-weakly supervised fashion with mask annotations and image-level labels l_i, a processing step is added where the pseudo-labels are filtered. Specifically, we integrate the image-level labels by adjusting a given pseudo-label $\hat{p}_i$ such that the classes occurring in the pseudo-label are also present in l_i. This process singles out some faulty segments in the pseudo-label, directly improving its quality. This naive and simple integration of weak labels is a first step towards assessing whether strong- and weak annotations can complement each other and produce an advantageous segmentation performance while giving rise to a more flexible annotation budget allocation.

4.2.3 Mean-taught deep supervision

Lastly, we propose to integrate a process which aims at producing more robust pseudo-labels through training towards invariance to perturbations in input space, *i.e.* when a single image is augmented differently, and towards consistency in output space, such that differently parameterized models lead to a similar segmentation. The *Self-taught Deep Supervision* variant will be extended with the paradigm of a mean-teacher setup [104], and thus, we refer to this alternate version as *Mean-taught Deep Supervision*. The idea of mean-teachers resides in keeping track of a second so called teacher network which shares the same architecture as the student network. Yet, the teacher is not updated by error-back-propagation, but rather by the exponential moving average of the student parameters over the previous iterations:

$$\theta_t^{\text{teacher}} = \alpha \cdot \theta_{t-1}^{\text{teacher}} + (1-\alpha) \cdot \theta_t^{\text{student}} \quad , \tag{II.11}$$

where $\theta_t^{teacher}$ are the teacher's parameters at iteration t, similarly $\theta_t^{student}$ are the parameters of the student at the same iteration and α is a smoothing coefficient. This formulation was previously introduced for semi-supervised classification [104] and was adapted for semantic segmentation [108, 109, 107] afterwards. By continuously

updating the teacher model with the parameters of the student, the moving average of those parameters is said to produce a model which is more robust, as it encompasses a combination of all models in previous iterations. Tarvainen *et al.* formed an objective function by aiming at aligning the student's softmax predictions to the teacher's via a mean-squared error (MSE) loss [104]. We extend this by also utilizing the teacher to derive hard pseudo-labels and leverage them in our *Multi-label Deep Supervision* that the student is trained with.

Previous adaptations of the mean-teacher framework to semantic segmentation did not perturb the input image to the teacher model [107] or made use of CutMix-like augmentations [108, 109]. What we propose is to differently perturb the input image to the teacher as opposed to the input to the student, but still obtain two outputs that align pixel-wise for successive pseudo-label supervision. Specifically, the input to the student network is an image which is augmented with commonly used augmentations (color jittering, flipping), the weakly augmented image x^{weakly}. The input to the teacher network applies further augmentations to x^{weakly}. Firstly, stronger photometric perturbations $\gamma(\cdot)$ are added (*e.g.* color jittering) and afterwards geometric augmentations $\varphi(\cdot)$ (*e.g.* flipping) leading to a strongly augmented version of the image $x^{strongly} = \varphi(\gamma(x^{weakly}))$. To achieve aligned outputs between the teacher and the student networks, after the forward pass of $x^{strongly}$ through the teacher network to obtain softmax predictions, the geometric augmentations are reversed, which we note down by $\varphi^{-1}(\cdot)$. As such, the aligned pseudo-label from the teacher's prediction is computed via Equation (II.10) and geometrically altered through $\varphi^{-1}(\hat{p})$. This process is an important detail to make the mean-teacher work for semantic segmentation and commonly used image augmentation strategies.

As the teacher network is said to produce more robust predictions, it is also used for inferring the semantic segmentation of test images. The complete *Mean-taught Deep Supervision* setup is displayed in Figure 6.

4.3 Experiments and results

In this section, we first describe the datasets and experimental setup to test the *Multi-label Deep Supervision* loss as well as the *Self-* and *Mean-taught Deep Supervision* training strategies. To rigorously test them, we outline our evaluation protocol for training runs with very few annotations as well as different annotation types and explain competing methods we compare to. Finally, we present the quantitative and qualitative results which lead us towards addressing our research questions as previously outlined (see Figure 3).

4.3.1 Datasets

To investigate scarce annotation training schedules and semi-weakly supervised semantic segmentation, we build our experiments on top of the expert-driven, medical domain of retinal fluid segmentation in optical coherence tomography (OCT) scans. Here, we obtain b-scans (2D images) from the imaged volumetric data from the RETOUCH dataset [8]. The volumes are distributed among three different OCT device types, Spectralis, Cirrus and Topcon, where we carry out the majority of experiments on the Spectralis b-scans, and verify the results on the data of Cirrus and Topcon on a smaller set of experiments. We chose the Spectralis vendor for the main experiments due to it being the smallest of the datasets (49 b-scans per volume vs. 128 b-scans per volume), which suits the setting of working with small amounts of data well. As the imaged b-scans from the three vendors differ in appearance quite significantly, as seen in Figure 7, we also do not consider training models on a larger fused dataset. This is in coherence with literature [8] and again suits our objective of working with small datasets and few annotations. The dataset is fully annotated with pixel-wise masks for three types of retinal fluids: *Intraretinal Fluid*, *Subretinal Fluid* and *Pigment Epithelial Detachments* for which we automatically derive bounding boxes and image-level labels to investigate weakly supervised as well as semi-weakly supervised segmentation scenarios.

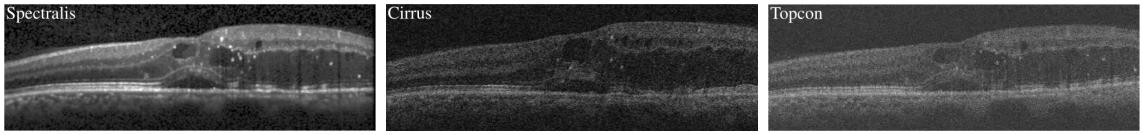


Figure 7: B-scan images taken from [8] indicating how diverse optical coherence tomography devices from different vendors image the approximately same region in the same patient's retina. Spectralis, Cirrus and Topcon devices produce different contrast, noise and image resolution.

4.3.2 Evaluation protocol

Our research questions encompass investigating the efficacy of semantic segmentation algorithms when faced with extremely few annotations, *i.e.* starting with merely one example per semantic class and successively doubling the amount. Thus, for the three classes in the RETOUCH dataset, we consider scenarios with 3, 6, 12 and 24 pixel-wise annotated b-scans. Therefore, we enumerate all training b-scans and further make sure, that in an interval of the size three, all diseases are present (as far as possible). With this setup, we can ensure, that in each of the 3, 6, 12 and 24 supervision scenarios at least 1, 2, 4 and 8 images contain each of the three classes. By enumerating the training set in this way, we also guarantee that *e.g.* the scenario with 24 annotated b-scans subsumes the scenarios with 3, 6 and 12 annotations. Put differently, we successively extend the small sets of annotated b-scans to investigate the effect of adding annotations in the early process of compiling datasets and how segmentation performance changes with it.

We will consider two main streams of experiments, which use (1) pixel-wise masks as the strongest type of annotations and (2) bounding boxes as the strongest annotation type. Both these streams of experiments are further divided into two flavours of experiments, where we investigate them either in a semi-supervised fashion, where we pair the masks or boxes with unlabeled b-scans or in a semi-weakly supervised fashion in which we investigate them paired with image-level labels. With these training and evaluation setups, we are able to answer our three research questions.

To end up at robust results, we carry out all the mentioned experiments with ten-fold cross-validation. This respects the sensitivity in the optimization process in low annotation scenarios and as we re-shuffle and re-enumerate the training images in each cross-validation split, the influence of specific b-scan annotations is reduced. The splits are generated randomly and independent for each vendor but stay the same when different algorithms are evaluated. The process is as follows: We randomly select 5 volumes for validation and 5 for testing, while the remaining volumes (Spectralis: 14, Cirrus: 14, Topcon: 12) belong to the training set. As we train on the b-scans which are extracted from those volumes and not on the volumes directly, with this setup we ensure cross-volume validation (*i.e.* the entire test and validation volumes are unseen during training and stem from entirely different patients).

To evaluate the efficacy, we follow the standard procedure for segmentation models and infer the class-prediction P for all pixels in all testing b-scans with a given model and calculate the Intersection over Union (IoU) using the ground-truth G:

$$\mathrm{IoU}(P, G) = \frac{P \cap G}{P \cup G} \tag{II.12}$$

To evaluate the performance on retinal fluid segmentation, we calculate the IoU for all C classes (*i.e.*, Intraretinal Fluid, Subretinal Fluid and Pigment Epithelial Detachments):

$$\mathrm{mIoU}(P, G) = \frac{1}{C} \sum_{c}^{C} \frac{P_c \cap G_c}{P_c \cup G_c} \tag{II.13}$$

The mean IoU is computed by averaging these individual class IoUs, here P_c and G_c refer to binary predictions and ground-truth related to class c. As we perform ten-fold cross-validation the final measures which we display are the average mean IoU and the standard deviation over these S splits (*i.e.*, each result reflects ten trained segmentation models):

$$\text{average mIoU} = \frac{1}{S} \sum_{s}^{S} \mathrm{mIoU}(P^s, G^s) \tag{II.14}$$

Here, P^s and G^s are the predictions and ground-truth of the current split s. In the following results we generally write mIoU as shorthand for average mean IoU.

4.3.3 Implementation details

Before training segmentation models on the individual b-scans extracted from the volumes, we apply a retinal OCT pre-processing [227], where anisotropic filtering is done and the lower edge of the retina is warped such that it approximates a straight line. The segmentation models are further trained on resized b-scans of size 200×200 enabling bigger mini-batches in training.

In order to produce directly comparable results, we employ the same semantic segmentation architecture for all our experiments. As the Unet architecture by Ronneberger *et al.* [44] has proven to be an off-the-shelf annotation-efficient segmentation variant as indicated by its successful application in many medical, expert-driven domains [8, 7, 228], we chose it as our backbone workhorse. Specifically, we implement all approaches on top of a fully convolutional Unet with batch norm layers [222] and four down-scaling encoder-blocks and four up-scaling (bi-linear interpolation variant) decoder-blocks. We refer to the feature maps after each of the convolutional decoder blocks as f_0 through f_4, where f_4 is the feature map directly before the output-head yielding full-size, pixel-wise predictions. For our *Self-* and *Mean-taught Deep Supervision* models, we utilize two output-heads on f_4 and for the *Multi-label Deep Supervision* loss individual output-heads on top of the feature maps $f_0, \ldots, f_3$ as specified in Section 4.1.2.

The main training hyper-parameters are tuned once for a vanilla Unet which utilizes mask annotations only in training. The mini-batch size is set to 16 b-scans, which are all augmented by horizontal flipping with a probability of 50% and a random adjustment of brightness, contrast, hue and saturation by a factor between 0.0 and 0.1. The network weights are initialized with Xavier initialization [229], the training is set to 100 epochs and stochastic gradient descent optimization with a momentum term of 0.9 is used. After 80 epochs the learning rate is adjusted from 0.01 to 0.001.

The best model is found via early stopping, *i.e.*, by evaluating models every 10 epochs on the validation set and using the best performing model on the validation set in terms of mIoU and apply it to the test set once after training is completed.

4.3.4 Competing approaches

Standard Unet: As all approaches use a similar network architecture, namely the encoder-decoder network Unet, our first baseline is simply training it using the strongest available annotations in the training setup with the cross-entropy loss of Equation (II.6). This means, if the strongest annotation type available is pixel-wise mask annotations, we only use those and do not consider unlabeled- or weakly labeled images. In case we have bounding boxes as strongest annotation type, we train the Unet similarly by considering the boxes as coarse pixel-wise masks. This is a naive weakly supervised integration of box annotations, yet it does not require designing hand-crafted constraints for different imaging domains. For small expert-driven domains such constraints generally have to be evaluated and potentially designed anew as standard approaches from natural image domains [166, 167] might not work due to wildly different image properties. As only considering images with either masks or boxes results in considerably small numbers of iterations per epoch (3,6,12 or 24 iterations) in our scenarios, we train this variant for 1000 epochs instead of the regular 100 epochs. These Unets are the lower bound, all other algorithms should outperform this training strategy.

Multi-label Deep Supervision: We investigate if the *Multi-label Deep Supervision* loss already brings about improvements when it is integrated into the above Unet training, where only annotated pixel-wise masks or boxes are leveraged. This baseline extends the above standard Unet training with our novel loss for all decoder feature maps $f_0, \ldots, f_4$.

Invariant Information Clustering (IIC): For semi-supervised learning, we consider Unets, which extend the standard Unet training with an additional self-supervised loss term based on the invariant information clustering (*IIC*) loss of Ji *et al.* [230]

which is integrated for all unlabeled images on the feature map f_4. By integrating unlabeled data aside the labeled data, this model functions as lower baseline for semi-supervised training strategies.

Multiple-Instance Learning (MIL): As a lower baseline for semi-weakly supervised algorithms with access to image-level labels, we leverage a Multiple-Instance Learning (*MIL*) segmentation model. This model is partially trained with pixel-wise masks or bounding boxes through cross-entropy and on top integrates image-level labels by average pooling the feature map f_4 in the spatial dimension and successively classifying the pooled feature. For this classification, we enforce binary cross-entropy for multi-label settings as in Equation (II.8) with the image-level label as the target.

Deeply Supervised IIC and Deeply Supervised MIL: As part of our idea is based on integrating deep supervision, we show the performance of the *IIC*- and *MIL*-model when the losses are integrated through deep supervision into all feature maps $f_0, \ldots, f_4$ of the decoder. These baselines give insight into deeply supervised training for semi- and semi-weakly supervised segmentation, we refer to them as *Deeply Supervised IIC* and *Deeply Supervised MIL*.

Perone and Cohen-Adad: From semi-supervised segmentation literature in the medical domain, we further test the consistency-based approach of *Perone and Cohen-Adad* [107] by re-implementing it. Strictly following their training scheme led to diverging models, thus, we modified it by using cross-entropy- instead of the DICE loss [67] and adapt their α parameter to 0.5 without a ramp up phase as well as a simple balanced loss weighting.

Self-taught Deep Supervision and Mean-taught Deep Supervision: Lastly, our proposed training schemes *Self-taught Deep Supervision* and *Mean-taught Deep Supervision* are both able to make use of masks or boxes aside unlabeled- or image-level labeled images. Therefore, they can be trained either semi- or semi-weakly by integrating the pseudo-labeling techniques as outlined in Section 4.2.2 and Section 4.2.3.

4.3.5 Hyper-parameter sensitivity studies

We investigate the performance of our *Mean-taught Deep Supervision* on the validation sets of the ten cross-validation splits in the scenario where only 24 annotated masks are available and the remaining images are supplied with image-level labels.

The experiments successively add different parts of our method, starting from not using a mean-teacher model but merely the network itself to provide pseudo-labels, *i.e.* starting from the *Self-taught Deep Supervision* model in the first line of Table 1. Then, in the second line, we add the teacher model with a smoothing coefficient $\alpha = 0.1$. Here, the standard mean-teacher MSE loss is integrated as well as our pseudo-label based *Multi-label Deep Supervision* loss which improve the performance slightly. Increasing the severity of photometic augmentations γ (brightness, hue, contrast, saturation) from a factor of 0.1 to 0.4 also increases the performance a bit in line three. Inferring segmentation results with the mean-teacher model instead of the student as well as increasing the smoothing coefficient to 0.5 lead to absolute improvements of +1.61% and further +1.21% in validation mIoU as indicated in lines four and five. The configuration in line five results in the best segmentation as indicated by the validation accuracy which is why we use it for the main experiments. The last line in Table 1 omits the MSE loss from the standard mean-teacher setup [104] to show, that without it, our *Mean-taught Deep Supervision* training strategy still works and is not dependent on it. For additional experiments which investigate the

larger γ	inference	α	MSE	validation mIoU
–	student	0.0	–	57.80 ± 4.68
–	student	0.1	✓	58.26 ± 4.27
✓	student	0.1	✓	58.54 ± 3.62
✓	teacher	0.1	✓	60.15 ± 4.14
✓	teacher	0.5	✓	$\mathbf{61.36 \pm 4.73}$
✓	teacher	0.5	–	61.24 ± 3.69

Table 1: Ablation for *Mean-Taught Deep Supervision* using 24 masks and image-level labels. First line indicates the *Self-Taught Deep Supervision* performance.

performance of different configurations of the baseline approaches *IIC*, *MIL* and their deeply supervised versions, please refer to Appendix A.

4.3.6 Quantitative results

After setting the hyper-parameters and validating that we have strong baselines for the scenarios of semi-supervised- and semi-weakly supervised training we now turn our attention towards the testing results of the ten splits for retinal fluid segmentation. First, in Table 2 we show the results when training segmentation algorithms with costly pixel-wise masks as strongest supervision. The columns indicate how many such masks were used to train the segmentation models ranging from merely 3, *i.e.* one mask per class, to 24 masks with the upper limit termed *Full Access*, where all 416 masks of the training split are used to train with. Two experiments make use of only these masks in training, the Unet baseline and the Unet trained with our *Multi-label Deep Supervision* loss in addition. As expected, the Unet baseline in the first row, which is the lower baseline for all models, improves for each scenario with more masks. Likewise, as expected, successively adding more masks comes with diminishing returns, the first three additional masks (3 to 6) come with a +82.3% relative improvement, while adding six more (6 to 12) increases the results by a relative +31.2% and the next added 12 masks (12 to 24) gives +37.4% relative improvements. When adding the *Multi-label Deep Supervision* loss to the same mask-only training, quite similar relative improvements are achieved when moving from 3 to 6 and 6 to 12 masks, with +83.1% and +30.5%, respectively. Yet, it already starts with a +3.18% higher mIoU in the 3 annotated mask scenario and therefore leads to a better segmentation accuracy throughout all scenarios. This improvement still holds when all masks are available, where our loss function improves the standard Unet performance by +3.73% mIoU showing that the multi-label loss on different scales in the decoder leads to finding better local minima in the optimization process. For an annotation-efficient training the desire is to have a higher starting performance with fewer mask annotations (*e.g.* at 3 masks) and simultaneously have the steepest

	Method	Mask Supervision 3	6	12	24	Full Access
	Baseline [44]	14.80 ± 6.50	26.98 ± 7.83	35.39 ±6.36	48.63 ±5.17	62.09 ±4.77
	Multi-label Deep Supervision (Ours)	**17.98 ± 8.20**	**32.92 ± 7.35**	**42.96 ±6.71**	**52.68 ±6.82**	**65.82 ±4.64**
Semi-sup.	IIC Baseline[8] [230]	**22.45 ± 9.36**	32.02 ± 7.23	41.48 ±7.26	53.08 ±6.13	65.16 ±3.80
Semi-sup.	Deeply supervised IIC[8]	20.78 ± 8.83	31.39 ±10.26	39.18 ±6.94	50.10 ±7.92	65.18 ±3.85
Semi-sup.	Perone and Cohen-Adad[10] [107]	16.17 ±10.74	33.10 ±10.24	45.80 ±7.51	54.75 ±5.96	65.49 ±4.14
Semi-sup.	Self-taught Deep Supervision (Ours)	10.37 ± 8.29	28.62 ±12.96	43.57 ±9.97	56.11 ±6.30	66.24 ±4.67
Semi-sup.	Mean-taught Deep Supervision[10] (Ours)	16.31 ±15.48	**35.17 ±11.35**	**53.52 ±8.72**	**58.84 ±6.57**	**66.31 ±4.66**
Semi-weak	MIL Baseline	15.44 ±11.10	25.46 ± 8.57	41.34 ±9.66	49.07 ±8.20	61.50 ±5.64
Semi-weak	Deeply supervised MIL	20.02 ± 9.17	31.50 ± 8.88	44.29 ±5.03	51.13 ±3.93	62.04 ±3.92
Semi-weak	Self-taught Deep Supervision (Ours)	20.47 ± 8.62	36.40 ± 8.91	49.39 ±9.95	59.29 ±7.52	66.34 ±3.81
Semi-weak	Mean-taught Deep Supervision[10] (Ours)	**21.91 ±13.49**	**42.14 ±14.25**	**54.70 ±9.26**	**60.45 ±5.71**	**66.39 ±4.29**

Table 2: Results on Spectralis in average mIoU over 10 splits with standard deviation for a set of algorithms trained with varying amounts of mask annotations (3, 6, 12, 24, all). We compare approaches using only masks, semi-supervised approaches adding unlabeled images and semi-weakly supervised algorithms utilizing also image-level labels (**best results in category bold**). Superscripts indicate smaller batch sizes.

relative increase in performance early on (*e.g.* from 3 to 6 masks).

Next, we look into the performance of semi-supervised segmentation algorithms highlighted with pink in Table 2. The multi-task Unet with the added IIC loss performs best in the extreme case with merely 3 mask annotations, where it achieves 22.45% mIoU. All approaches which construct pseudo-targets from the unlabeled data and integrate them with a loss function, namely *Perone and Cohen-Adad* and our *Self-taught-* and *Mean-taught Deep Supervision* methods struggle in this extreme case and in the case of *Self-taught Deep Supervision* even lead to a degradation below the standard Unet results. This behaviour can be explained by the known problem of confirmation bias [231] which leads to overfitting to the faulty pseudo-targets when the predictions are unreliable. With more expert-annotated masks, *i.e.* 6, 12 and 24, these methods are better equipped to approximate the unavailable ground-truth for the unlabeled b-scans and more reliably outperform the IIC multi-task method. In

these three scenarios the design choices which lead from the *Self-taught-* to the *Mean-taught* variant of our method show to be critical. The move to include the mean-teacher structure and the training towards higher consistency between differently perturbed versions of the same image for pseudo-label computation in the *Multi-label Deep Supervision* loss result in improvements of +6.55%, +9.95% and +2.73% mIoU. With these results, the *Mean-taught Deep Supervision* method achieves the best mIoU values for the scenarios from 6 masks to full access in the semi-supervision category. Something to note is, that both the self-supervised proxy-task of IIC and the consistency loss terms of the remaining semi-supervised methods also led to improvements in the setting when all b-scans have associated masks available.

In Section 4.2.2 we outlined a process to integrate image-level labels into the training procedure of the *Self-taught-* and *Mean-taught Deep Supervision* models. This semi-weakly supervised training of our proposed methods is compared to the Multiple-Instance Learning baselines highlighted with cyan in Table 2. The integration of image-level labels via Multiple-Instance Learning in the *MIL baseline* provides a weak learning signal which only starts to consistently outperform the standard Unet trained solely with masks when integrating it into all layers of the decoder, *i.e.* with the *Deeply supervised MIL* variant. Yet, in the case when all masks are available, this semi-weakly training strategy does not improve upon the standard Unet basline, as the semantic information from image-level labels do not add to the semantic masks which are present for all b-scans, while the consistency training in our approaches still help. Where image-level labels clearly help is in counteracting the confirmation bias of our pseudo-label-based *Self-taught-* and *Mean-taught Deep Supervision* methods early on in the 3 mask +413 image-level label scenario and the 6 mask +410 image-level label scenario. There, they lead to improvements of +10.1% and +7.78% mIoU for the *Self-taught Deep Supervision* method and to +5.6%, +6.97% for *Mean-taught Deep Supervision*. The large increase in absolute mIoU in the starting scenario of 3 mask +413 image-level labels is coupled with a steep relative increase of +92.3% when adding 3 more mask annotations, showing the annotation-efficiency of the

	Bounding Box Supervision				
Method	3	6	12	24	Full Access
Baseline [44]	12.49 ±4.28	18.32 ±4.94	25.62 ±3.08	29.55 ±2.77	38.45 ±4.44
Multi-label Deep Supervision (Ours)	**14.59 ±5.81**	**19.62 ±6.21**	**27.89 ±3.44**	**32.02 ±4.78**	**38.66 ±3.36**
Semi-sup. IIC Baseline[8] [230]	**15.40 ±7.07**	18.15 ±7.49	26.05 ±6.00	30.07 ±4.32	38.45 ±4.65
Semi-sup. Deeply supervised IIC[8]	12.77 ±7.15	17.76 ±6.26	**28.99 ±4.60**	30.64 ±3.05	38.81 ±4.48
Semi-sup. Perone and Cohen-Adad[10] [107]	11.17 ±7.41	**19.02 ±8.46**	27.44 ±5.81	31.72 ±3.87	39.38 ±3.56
Semi-sup. Self-taught Deep Supervision (Ours)	5.14 ±3.84	9.62 ±7.35	24.47 ±6.12	32.71 ±3.56	**39.39 ±3.63**
Semi-sup. Mean-taught Deep Supervision[10] (Ours)	8.21 ±3.96	14.28 ±7.48	24.79 ±5.79	**34.14 ±3.10**	39.04 ±4.15
Semi-weak MIL Baseline	15.82 ±6.55	16.95 ±6.19	22.56 ±4.56	26.48 ±5.51	37.15 ±4.06
Semi-weak Deeply supervised MIL	**17.14 ±8.06**	20.18 ±4.61	24.15 ±4.95	29.12 ±4.75	37.94 ±3.35
Semi-weak Self-taught Deep Supervision (Ours)	16.04 ±8.52	**22.15 ±6.29**	28.63 ±4.04	32.37 ±3.75	**38.97 ±3.59**
Semi-weak Mean-taught Deep Supervision[10] (Ours)	15.81 ±8.59	21.97 ±8.17	**29.83 ±5.30**	**34.81 ±3.62**	38.66 ±4.73

Table 3: Results on Spectralis in average mIoU over 10 splits with standard deviation for a set of algorithms trained with varying amounts of box annotations (3, 6, 12, 24, all). We compare approaches using only boxes, semi-supervised approaches adding unlabeled images and semi-weakly supervised algorithms utilizing also image-level labels (**best results in category bold**). Superscripts indicate smaller batch sizes.

Mean-taught Deep Supervision method.

In Table 3, we carry out the same experiments as in Table 2 but exchange the mask annotations with cheaper bounding box annotations as strongest supervisory signal. The observation which stays the same in this setup is that integrating the *Multi-label Deep Supervision* loss again improves the results as opposed to standard Unet training, although more gradually. This behaviour can be traced back to the much lower performance ceiling of 38.45% mIoU of a Unet trained with bounding boxes for all b-scans. While our approaches still perform favourably in the semi-weakly supervised category for scenarios with 6 masks and more, the margins are generally much less pronounced. To compare the annotation types used in Table 2 and in Table 3, one observation that can be made is that the models trained with full access to 416 bounding boxes still performs worse than semi-supervised models with access to merely 12 masks, and worse than semi-weakly supervised models using

	Method	Mask Supervision 6	12	24	Full Access
	Cirrus				
	Baseline [44]	12.31 ±5.41	19.43 ±8.00	30.10 ± 9.34	48.92 ±11.94
	Multi-label Deep Supervision (Ours)	**15.99 ±6.87**	**25.12 ±8.58**	33.53 ± 9.44	50.47 ±10.84
Semi	*Perone and Cohen-Adad*[10] [107]	12.36 ±6.12	24.99 ±6.49	33.79 ±10.15	49.75 ±12.87
Semi	Mean-taught Deep Supervision[10] (Ours)	9.18 ±8.53	23.33 ±7.37	**35.82 ±11.40**	**51.24 ±10.94**
	Topcon				
	Baseline [44]	14.79 ± 9.34	21.19 ±11.57	27.61 ±10.31	42.22 ±10.42
	Multi-label Deep Supervision (Ours)	**18.20 ±10.48**	20.92 ±13.02	33.71 ±11.92	**45.85 ±10.32**
Semi	*Perone and Cohen-Adad*[10] [107]	15.26 ±12.74	21.88 ±12.48	27.67 ±13.81	41.43 ± 8.18
Semi	Mean-taught Deep Supervision[10] (Ours)	14.39 ±11.19	**23.92 ±15.25**	**33.87 ± 8.25**	42.70 ±10.97

Table 4: Results in average mIoU over 10 splits with standard deviation for a set of algorithms trained with varying amounts of mask annotations (6, 12, 24, all) on data of OCT vendors Cirrus and Topcon. We compare approaches using only masks and semi-supervised approaches adding unlabeled images (**overall best results bold**). Superscripts indicate smaller batch sizes.

only 6 masks and 410 cheaper image-level labels. Thus, an annotation of an image with bounding boxes needs to be obtained in about $\frac{12}{416} \approx 2.88\%$ of the time of one annotated pixel-wise mask to be economical, when the target performance lies at around 40% mIoU. As the target performance is set higher, with the diminishing returns of additional annotations and the large gap in best performance between mask- and box annotated images, a target performance of *e.g.* 60% mIoU would require much more box annotated images than are in the training set. On such a small dataset the semi-weakly supervised paradigm with our *Mean-taught Deep Supervision* is still able to achieve this performance with merely 24 mask annotations and 392 image-level labels.

In Table 4, we expand some of the experiments to the Cirrus and Topcon datasets within RETOUCH [8] to get a broader picture. There, our *Multi-label Deep Supervision* loss and the *Mean-taught Deep Supervision* setup perform best. The semi-supervised scenarios with merely 6 pixel-wise annotated masks exhibit worse mIoU

scores as opposed to only training with masks, which might hint at a too stark influence of the pseudo-label-based loss terms due to a higher amount of unlabeled b-scans in the Cirrus and Topcon sets. This behaviour might further hint at confirmation bias from too imprecise predictions, which for these datasets is reduced when moving towards the 24 mask cases, *i.e.* when enough ground-truth masks are available to infer better pseudo-labels.

4.3.7 Qualitative results

For a visual display of the performance of different segmentation algorithms on retinal fluid segmentation, in Figure 8 we see how they match up against each other when successively supplied with more and more costly pixel-wise annotations. The top two rows display the segmentation results of algorithms trained with mask annotations only, the four rows below show the semi-weakly trained algorithms which are trained with masks and image-level labels. Comparing the Baseline Unet and *Multi-label Deep Supervision*, it can be observed that throughout the low supervision scenarios, the former produces small speckles in the segmentation, confusing different fluid types with one another, *i.e.* **Subretinal fluid** is segmented as **Intraretinal Fluid**. Concerning the semi-weakly supervised algorithms, both our *Self-taught-* and *Mean-taught Deep Supervision* methods are able to, with 6 masks only, localize the most significant regions where the three fluid types occur, even the **Pigment Epithelial Detachments** which is the most challenging class to segment in the dataset. The baseline methods struggle with segmenting the large **Subretinal fluid** regions up to the 24 mask scenario, where they start to capture it more fully.

4.4 Discussion

Medical applications for semantic segmentation systems are prime examples of expert-driven domains, where annotations are costly to acquire and largely depend on the availability of experts, *i.e.* medical doctors. To cope with these scenarios and still successfully train segmentation models with severely limited pixel-wise annotations, we

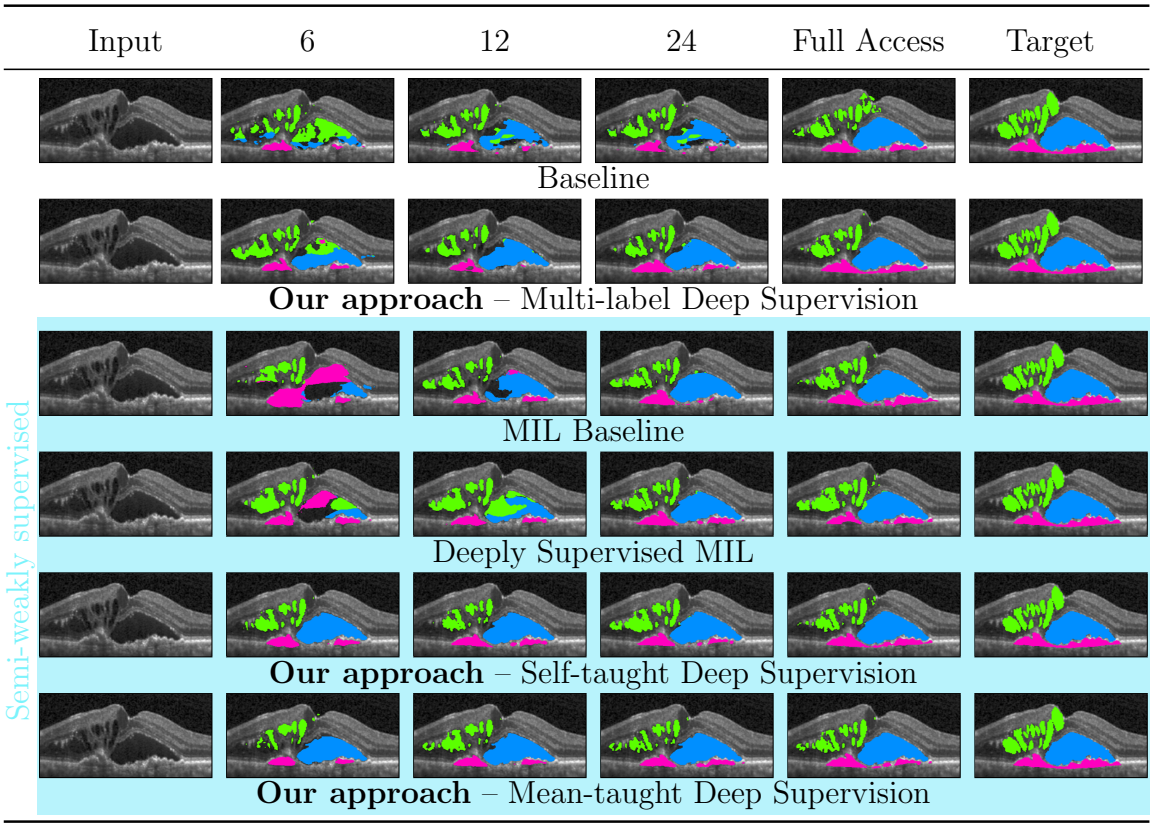


Figure 8: The progression of segmentation results for algorithms trained with successively more mask annotations, from 6 to 12, 24, and all available masks. The first two approaches use only masks, while the remaining four approaches are trained in the semi-weakly supervised style using masks and image-level labeled data.

addressed three research questions. We trained segmentation models with extremely few pixel-wise annotations and investigated how their performance progresses when additional masks are provided. In this respect, we saw, that segmentation models struggle considerably in the extreme case when only singular annotated images are available per class.

Further, we quantified, that every additional annotation comes with diminishing returns in performance, meaning that a similar increase in segmentation performance at low performance levels is associated with less annotation effort as compared to an

increase starting at a higher performance. By integrating unlabeled data using semi-supervised algorithms from literature, we are able to see, that with this learning paradigm, segmentation algorithms can achieve similar segmentation results with fewer costly pixel-wise annotations in the expert-driven domain at hand. We further, observed that with our *Multi-label Deep Supervision* loss and the flexible *Mean-taught Deep Supervision* method we are able to achieve a higher performance increase at early phases of annotation, where only very few, *e.g.* two annotations per class are available.

We also investigated the performance of segmentation algorithms when they are supplied with bounding boxes as strongest supervision signal instead of masks, in order to investigate whether the effort of obtaining costly pixel-wise annotations is worth it. In our investigations, we set up the segmentation algorithms to learn from bounding box supervision in a very simple way by exchanging the pixel-wise annotations with boxes. This regimen does not use hand-crafted additional constraints frequently used in weakly supervised literature, as they are often designed for natural imagery and might not be applicable to arbitrary expert-driven domains with imagery far from object-centric data. Thus, by looking into the performance of these simple weakly supervised algorithms, we were able to compare them with mask-supervised segmentation algorithms. Specifically, we see, that very few pixel-wise annotations, *e.g.* 12 masks are already able to outperform segmentation algorithms supervised by multiple hundreds of images annotated with bounding boxes. With the diminishing returns of additional annotations, to train algorithms sufficiently with high target performance goals, only resorting to box annotations is not necessarily more economical as masks are more expensive but far fewer mask annotations are necessary as opposed to box-annotated images. If the datasets are very small, completely annotating the images with boxes might fail in satisfying performance requirements, while our experiments show, that the flexible learning paradigm of semi-weakly supervised learning can fulfill them. This more flexible view on the algorithmic side of learning semantic segmentation models, *i.e.* being able to profit from different annotation

types has effects on the annotation process as well: narrow annotation processes that restrict themselves to one annotation type might not be optimal as each annotation comes with diminishing returns and the added semantic information, *e.g.* the wrongly assigned class for a segment in a specific image, might also be fixed by a coarser and cheaper image-level annotation. Especially when the annotation budget is small, being able to gradually narrow down a performance target with added coarse annotations spares the expert's time as opposed to mindlessly adding pixel-wise annotations. For imaging data that occupies a certain visual redundancy, such as the redundancy between b-scans that lie adjacent to each other in the original volume, adding pixel-wise annotations might even be unnecessary to a certain extent, which we investigate in the next chapter. Contributions of this chapter summarize as:

Contribution 1:
Current segmentation algorithms progressively get better in segmenting images when more and more annotations are supplied. With our *Multi-label Deep Supervision* loss, we increase the starting performance of models that are exposed to merely pixel-wise annotated data. Thus, models trained with this loss are able to get more out of the individual mask annotations by using them to generate a stronger learning signal and engraving semantic information already in earlier layers of the network.

Contribution 2:
The performance of models using costly pixel-wise masks is compared to models supplied with bounding boxes in the training process. This enabled the comparison of the number of mask annotations and bounding boxes needed to achieve varying target performances and sheds light on the efficacy of different annotation types for small datasets. Resorting to only bounding boxes entails a lower ceiling performance at the cost of annotating more images, which directly influences the annotation process where the question of annotating few masks or a lot of bounding boxes has to stand in relation.

Contribution 3:
We proposed the *Mean-taught Deep Supervision* method which is able to flexibly integrate different annotation modalities, such as pixel-wise annotated data or bounding boxes as well as image-level labels or unlabeled data. With our design, the annotation process can be broken up by not requiring a single annotation type but enabling the training with a mix of annotation types, which enables a more flexible and economic spending of annotation budgets.

While the exploration in this chapter already helps in making segmentation algorithms more annotation efficient, another angle that might enable better usage of the expert's availability for annotations might be to respect that some image data includes a lot of redundant visual information. To cope with this redundancy and not spend the expert's time on annotating redundant portions which only exhibit limited performance gains, in the next chapter, we investigate how to learn from volumetric data, when it is only partially annotated. Developing a training strategy that is able to learn from volumetric data without having to provide volumetric annotations could drastically reduce the annotation effort and make 3D volumetric segmentation better feasible in expert-driven domains.

5 Learning with partial annotations

This chapter opens up a pathway to circumvent expensive and prohibitively time-consuming dense volume annotation needed for volume segmentation training by designing a training algorithm which builds on partial annotations and entirely unlabeled volumes. We tackle this semi-weakly supervised volume segmentation scenario via designing a positional- and a semantic coherence constraint which we enforce through an auxiliary loss function which shapes a voxel-embedding branch. By testing our so called Contrastive Constrained Regularization (Con2R) approach and compare it to traditional semi-supervised volume segmentation algorithms on two medical volume segmentation datasets, we are able to show that it achieves the best performance. Specifically, Con2R is able to, with less than 4% of labeled sub-regions, still reach up to 88% segmentation accuracy as compared to a fully supervised baseline which has access to dense volumetric annotations.

This section is based on a publication in *ECCV 2022* [34].

5.1 Introduction

Apart from different appearance properties such as contrast, noise and a different amount of spectral channels, there often exists a difference in the imaged spatial dimensions in expert-driven domains. As such, many imaging procedures include an additional spatial dimension, making the imaged data volumetric in nature [221, 232, 233, 234, 235, 236]. Examples of volume data is often found in the medical domain, there computed tomography [232, 233], magnetic resonance tomography [234] or optical coherence tomography [221] can yield volumes, not merely individual images. In order to make use of this additional spatial dimension, researchers developed 3D segmentation algorithms [66]. While these volumetric segmentation algorithms have been proven successful, they are generally trained using fully annotated volumes. For applications where the expert's time to annotate is

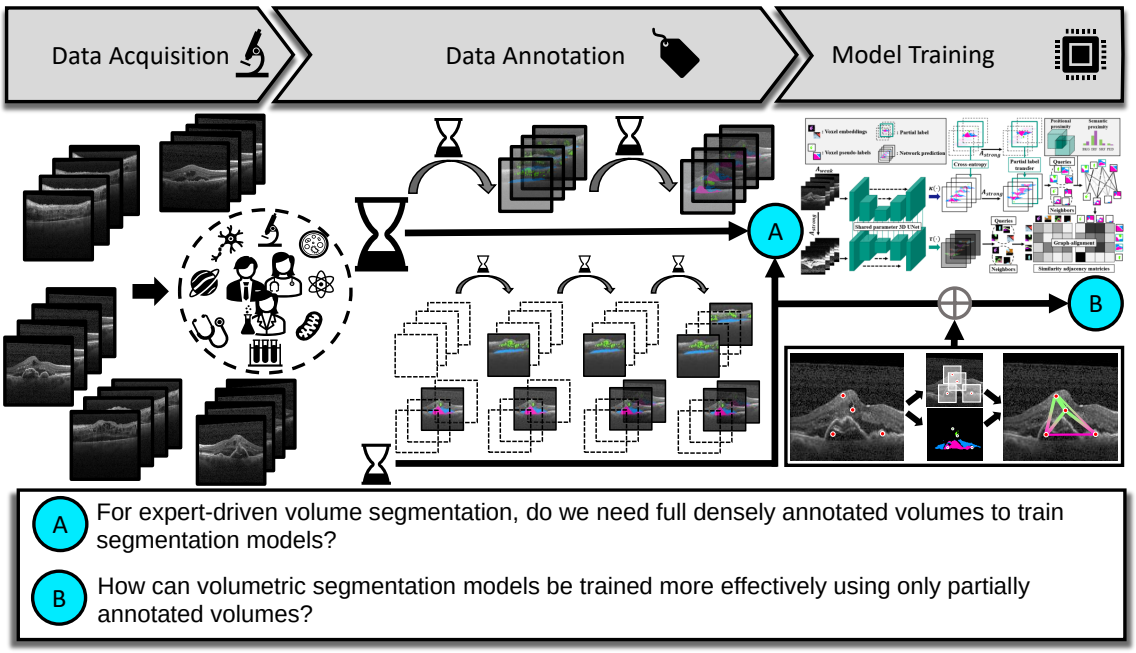


Figure 9: Overview of the main research questions in this chapter, they will be explored on an optical coherence tomography dataset [8] and magnetic resonance imaging [7] where medical doctors are needed in the annotation process. We further outline the *Contrastive Constrained Regularization* method which can help in expert-driven domains where volumetric data has to be segmented.

already scarce, requesting 3D annotations instead of 2D annotations further reduces the number of annotated samples proportional to the extent of the added dimension. Volumetric segmentation models, through the three dimensional input and three dimensional processing steps, *e.g.* 3D convolutions, have the potential to learn more sophisticated patterns useful for the segmentation task. Yet, the process of annotating full volumes can be put into question, as adjacent slices within the volumes often contain quite similar visual patterns, which might lead to redundancy in the annotation process and opens the question whether a different modus operandi for annotation might be more considerate towards the expert's availability.

In this chapter we want to start gathering insight into whether 3D segmentation models can be supervised in a more economical fashion by supplying them only with

partially annotated volumes. There, we are interested in how well a standard model can be trained with partial labels, with respect to its segmentation performance. To take into account semi-supervised literature, which generally utilizes full densely annotated volumes aside completely unlabeled ones, we perform semi-weakly supervised volume segmentation, where we make use of a mix of partially annotated volumes and unlabeled volumes in training. We develop a method for semi-weakly semantic volume segmentation by considering how to bring the three dimensional nature of the input image to the output predictions of a 3D segmentation model even though we only have partial, 2D label fragments. All models are tested on retinal fluid segmentation in optical coherence tomography [8] and brain tumor segmentation in magnetic resonance imaging [7].

We start by introducing the problem formulation of semi-weakly volume segmentation using partial and unlabeled data in Sec. 5.1.1. Then we present a solution to this training scenario, the *Contrastive Constrained Regularization* method, which encompasses a data-driven proxy task to address the issue of only having two dimensional targets while desiring three dimensional predictions. The method is motivated by the implications on the annotation process, which can be carried out in a more expert-centric fashion, potentially circumventing low-yield, redundant annotations and leaving the choice of interesting regions to annotate to the expert annotator. The research questions of this chapter are summarized visually in Figure 9.

5.1.1 Problem statement

We are concerned with the problem of segmenting volumetric data, *i.e.* 3D imaging data voxel-wise. Therefore, for any given input volume $v \in \mathbb{R}^{c_{dim} \times D \times H \times W}$, with c_{dim} input channels and the spatial dimensions depth D, height H, width W of the volume to be segmented, we want to predict the class association of each voxel forming the prediction $p \in \{0,1\}^{C \times D \times H \times W}$ with C class indices to segment.

In standard supervised training, every volume in the training set has to be associated with a complete dense mask annotation $m \in \{0,1\}^{C \times D \times H \times W}$, which is prohibitively

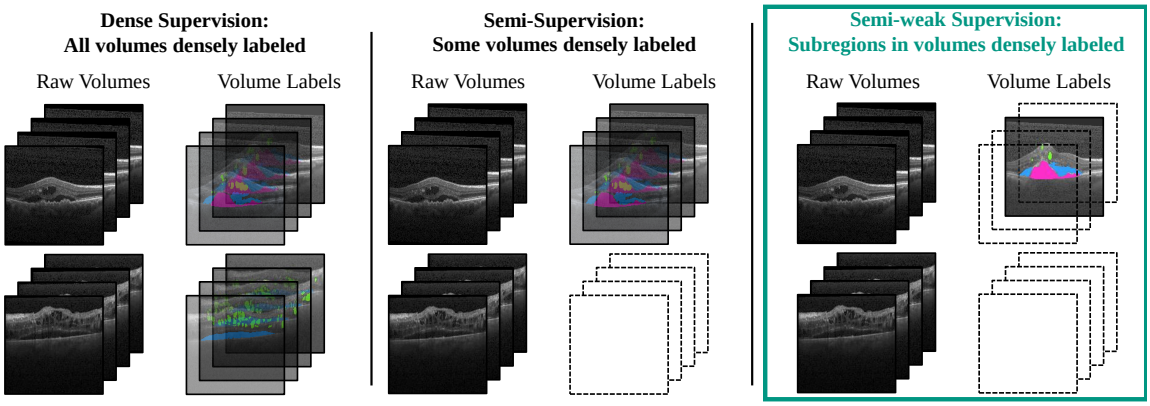


Figure 10: *Left*: densely supervised volume segmentation builds upon fully annotated volumes which are prohibitively expensive and might suffer from redundancy in adjacent slice annotations. *Center*: semi-supervision allows volumes without annotations, redundancy issue remains. *Right*: we propose to train with partially- and unlabeled volumes, freeing up experts to annotate across different volumes providing diverse annotations, while lowering the risk of spending time on redundant/uninformative volume portions.

expensive and thereby shrinks the deliverable amount of samples in a dataset. Data-efficient learning for semantic volume segmentation is generally carried out through a semi-supervised paradigm, where algorithms are supplied with a few fully annotated volumes and completely unlabeled volumes. What we investigate is the case, when we do not want to burden the expert annotators with annotating complete volumes but just regions within a volume. Therefore, instead, we train from partially labeled volumes, where only a subset of 2D slices of the volumes are annotated, while the rest of the volume is unlabeled. Furthermore, we pair those partially labeled volumes with completely unlabeled volumes. This semi-weakly training strategy as well as the other two learning paradigms are displayed in Figure 10.

To formalize this semi-weakly volume segmentation, we first define a dataset of volumes:

$$\mathcal{D} = \{v_1, ..., v_N | v_i \in \mathbb{R}^{c_{dim} \times D \times H \times W}\} \ , \tag{II.15}$$

containing a total number of N volumes. In our training setting, instead of having

complete masks m associated to each of the volumes in $\mathcal{D}$, we train models with access to:

$$\mathcal{M} = \{(m_1, a_1), ..., (m_N, a_N) | (m_i, a_i) \in (\mathbb{R}^{C \times D \times H \times W}, \{0,1\}^D)\} \ . \tag{II.16}$$

This setting defines for a volume $v_i \in \mathcal{D}$ both a ground-truth mask m_i and a binary indication $a_i^d \in \{0,1\}$ which indicates whether at the depth position $1 \leq d \leq D$ a slice annotation is present or not. In case $a_i \doteq 0^D$, the volume v_i is completely unlabeled, in case it contains D ones, it has a label for each voxel in the volume.

We are interested in data-scarce scenarios, therefore, we conduct experiments mostly in settings where $\sum_{d=1}^{D} a_i^d \ll D$, *i.e.* volumes v_i are only partially and very sparsely annotated. As this still allows for a high amount of partially annotated volumes, which might be out of reach for expert-driven scenarios, we also define, that in total, we only have access to a small number of annotated slices in the whole dataset: $\sum_{i,d=1}^{N,D} a_i^d \ll N \cdot D$. With the training data $\mathcal{D}$ and $\mathcal{M}$ we still want to train a volume segmentation model, which is able to, even though trained with partially annotated volumes, produce full volumetric segmentations for new, never seen before volume data. While the goal stays consistent with fully supervised segmentation, the alteration of the training scenario enables the experts to also annotate partial volumes and use their time to cover more, visually diverse volumes rather labeling very similar, volumetrically adjacent regions.

5.1.2 Preliminaries

For easier understanding, next the notation for volume indexing and volume processing is defined. We index an input volume v, or any volumetric tensor for that matter, using a three dimensional voxel location x by writing v^x. The interest in our investigation lies in volume segmentation architectures such as 3D encoder-decoder models [66], which we formalize to produce voxel-wise features $f \in \mathbb{R}^{f_{dim} \times D \times H \times W}$. These features f can be turned into semantic predictions p by using an output-head

$\kappa(\cdot)$ which is parameterized by C $1 \times 1 \times 1$ convolution kernels and trained using the few annotated voxels in the partially labeled volumes. Additionally, to this standard 3D processing pipeline, our method makes use of an additional transformation head which we note down as $\tau(\cdot)$ and parameterize by a normalization layer, $1 \times 1 \times 1$ convolution layer, a LeakyReLU activation function and a final e_{dim} $1 \times 1 \times 1$ convolutional kernels. Similarly to $\kappa(\cdot)$, $\tau(\cdot)$ operates on the voxel-wise features f to produce voxel-wise embeddings $e \in \mathbb{R}^{e_{dim} \times D \times H \times W}$. An embedding e^x at location x is meant to describe the imaging properties of the input voxel v^x in terms of an e_{dim} dimensional vector, while the corresponding semantic prediction p^x captures the class association at the same input voxel. Next, we describe our method, *Contrastive Constrained Regularization (Con2R)*, for training volume segmentation models with only sparse and unlabeled data.

5.2 Graph-constraints as regularization

In our semi-weakly supervised volume segmentation scenario we are not supplied with densely labeled volumes, but only have access to sparsely- or not at all labeled data. Therefore, we opt to design data-driven constraints on the hypothesis space of valid model parameter configurations. These constraints shall be designed in such a way as to nudge the model towards a configuration which, in the output space, where we are only supplied with sparse annotations, nonetheless produces three dimensionally coherent predictions. To achieve this, we take the common view of the input volume as a graph [135, 136, 133, 137]. Specifically, we design a complete bi-partite weighted graph $\mathcal{G} = (\mathcal{Q}, \mathcal{N}, \mathcal{E}, \sigma)$. This graph consists of two sets of vertices, which we term the Query-set $\mathcal{Q}$ and the Neighborhood-set $\mathcal{N}$, both of which containing voxel-embeddings e^x produced by the transformation head $\tau(\cdot)$ and sub-sampled[1] from all locations in the volumetric voxel-embedding tensor e. For our

[1]The sub-sampling step and design as a bi-partite graph is introduced to reduce the computational demand, a graph where all voxels are fully connected to each other is computationally infeasible, especially in our case where we intend to compute pairwise similarities and backpropagate through each connection.

purposes, we set the size of the two sets of voxel-embeddings $|\mathcal{Q}| = |\mathcal{N}|$, though, this can be chosen differently. In the graph $\mathcal{G}$, the vertices e^x in the Query-set $\mathcal{Q}$ are connected to all embeddings $e^y \in \mathcal{N}$ through edges $(x, y) \in \mathcal{E}$. Importantly, these edges carry a weight, which we set to $\sigma(x, y) = e^{xT} e^y / (||e^x|| \cdot ||e^y||)$, *i.e.* the cosine similarity between the voxel-embeddings they connect, making $\mathcal{G}$ a similarity graph. With this formulation, it is now possible to quantify how different voxel-embeddings, or, the input voxels they encode, relate to each other in the currently trained model. As the computation of the similarity graph $\mathcal{G}$ is differentiable, we can enforce the model to align the current similarities of voxel-embeddings to some specified target similarities. This mechanism enables us to define similarities between individual voxel-pairs in order to regularize the model in the training process. At this point, we need to define what similarities two voxel-embeddings actually should have. For this, we need to design a function $\mathcal{T}(\cdot)$ which, for each voxel-embedding pair returns a specific target similarity, that the model should produce.

Let's consider the hypothetical case where we have dense supervision. In this case, choosing $\mathcal{T}(\cdot)$ to be solely based on the agreement between the annotation of the two vertices is a reasonable choice. This would enforce that voxel-embeddings belonging to the same class to be embedded similarly. In our experimental design, we do not have such dense annotations, therefore we need to base the target similarities on workable assumptions that augment the incomplete knowledge about the voxel semantics we have.

5.2.1 Graph-based contrastive constraints

Here, we describe the considerations that go into designing the target similarity function $\mathcal{T}(\cdot)$ and the subsequently derived data-driven constraints that we use to define it: the *receptive smoothness* and *semantic coherence constraints*.

Receptive smoothness constraint In semi-supervised learning, there is the well-known smoothness assumption [237] stating that *samples close to each other likely share a class label*. Prior work on semi-supervised learning [108, 104] used this

assumption to enforce consistent predictions between a single sample which was augmented in different ways. In a related fashion, we want to enhance this smoothness assumption to consider the magnitude of such a perturbation and formulate: *samples closer to each other are more likely to share a class label*. Taking the graph design introduced above, this assumption can be integrated into model training, by designing similarity targets which are conditioned on the magnitude of a perturbation. To this intent, we consider translations as a form of perturbation, and with this, we can specify the similarity between two voxel-embeddings to be proportional to the relative position shift of the corresponding voxels in the input volume. Considering the two embeddings $e^x \in \mathcal{Q}$ and $e^y \in \mathcal{N}$ we compose the *positional proximity* of two voxel-embeddings in the volume by using the relative intersection of sub-volumes centered at the positions x and y, which we smooth by a small ε if the intersection approaches zero:

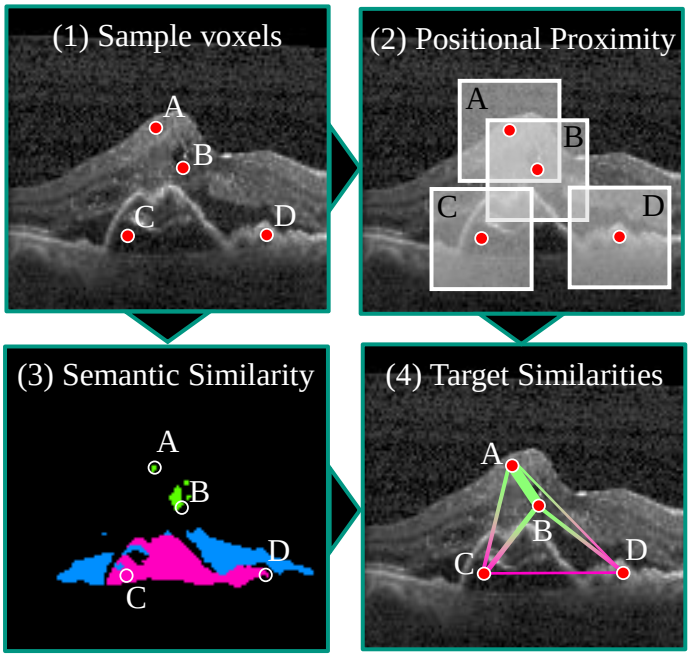


Figure 11: Simplified 2D graph. We impose constraints on the relationships between pairs of voxels. These constraints are determined by positional proximity, which is assessed based on the overlap of receptive fields, and similarity in class predictions.

$$\rho(x, y, \mathcal{R}(\cdot)) = \max\left(\frac{|\mathcal{R}(x) \cap \mathcal{R}(y)|}{|\mathcal{R}(x)|}, \varepsilon\right) \ , \tag{II.17}$$

where the receptive field function $\mathcal{R}(\cdot)$ returns for a voxel position x all spatially related voxels that fall into the sub-volume centered at x. To display Equation (II.17)

visually, we depict this idea of quantifying how close two voxels are based on what portion of their receptive fields (or receptive volumes in our case) they share, in Fig. 11. For a more convenient visualization, we show the positional proximity idea for the 2D case, in the image labeled by (2). Our voxel-embeddings result from processing input volumes with a volumetric segmentation model, which in our case is an encoder-decoder architecture [66], and therefore, each voxel-embedding was computed via successively encoding and then decoding local neighborhoods in a convolutional fashion. Thus, for each voxel-embedding a defined amount of the input voxels went into computing it: the input voxels which fall into its receptive volume. By defining the positional proximity of two voxels through the portion of voxels they share in their receptive volume the notion of similarity has a direct link to their shared patterns in the input. Voxel B in Fig. 11 gets attributed a higher similarity to voxel A as opposed to voxel D, as it shares a larger portion of voxels that went into computing their embeddings with A and therefore should be embedded *closer* to A than to D. We form the positional proximity constraint $\mathcal{P}(\cdot)$ by marginalizing positional similarities $\rho(\cdot)$ over the whole neighborhood embeddings $e^z \in \mathcal{N}$:

$$\mathcal{P}(x, y, \mathcal{R}(\cdot)) = \frac{\rho(x, y, \mathcal{R}(\cdot))}{\sum_{e^z \in \mathcal{N}} \rho(x, z, \mathcal{R}(\cdot))} \quad . \tag{II.18}$$

In case we do not have any information about the class membership of voxels, this design of computing target similarities for the weights in $\mathcal{G}$ leads to the model having to encode the full extent of the three dimensional receptive volumes of the input into the embeddings from which the segmentation output-head $\kappa(\cdot)$ can profit. When we have only access to partial annotations, this might help in aligning an unlabeled voxel-embedding to the embedding of a spatially close labeled voxel in the embedding space and thereby enable $\kappa(\cdot)$ to better assign the matching semantic class.

Semantic coherence constraint To embed spatially close voxels more similarly than spatially far apart ones is sensible to encourage the formation of clusters based on shared local visual patterns, but does not fully respect the semantic segmentation

task. In semantic segmentation, the same semantic class can occur at far apart locations which using the positional proximity constraint in isolation, would lead to voxel-embeddings of the same class which lie far apart to be embedded maximally dissimilar. This would be the case in Figure 11 (3), where voxel C lies closer to B than to D, but actually shares the predicted semantics (pink class) with D, hence embeddings of C and D should be *more similar* than C and B. This shows, while useful, the positional proximity constraint has to be offset with a constraint that considers the possibility of two distant voxels belonging to the same semantic class and thus, enforce a coherence between them. For the embeddings e^x and e^y belonging to $\mathcal{Q}$ and $\mathcal{N}$, respectively, we consider the semantic predictions p^x and p^y generated by the segmentation output-head $\kappa(\cdot)$, which itself is trained using the few partial volume labels. Now, in order to quantify the semantic similarity $\mathcal{S}(\cdot)$ of two voxels at locations x and y using the softmax class prediction p at those locations, different functions have been proposed [136]. We apply the symmetrized negative Kullbalk-Leibler divergence as similarity measure:

$$\text{SN-KL}(p^x, p^y) = -\frac{1}{2} \cdot \left(p^y \cdot \log\left(\frac{p^y}{p^x}\right) + p^x \cdot \log\left(\frac{p^x}{p^y}\right)\right) \quad , \tag{II.19}$$

as for the positional proximity constraint, we also marginalize over the predictions at all locations of the neigborhood set:

$$\mathcal{S}(x, y, p) = \frac{\exp(\text{SN-KL}(p^x, p^y))}{\sum_{e^z \in \mathcal{N}} \exp(\text{SN-KL}(p^x, p^z))} \quad . \tag{II.20}$$

The continuous values of p^x and p^y can thereby be transformed into a single value, which is higher the more similar these semantic voxel-predictions are and smaller if the class predictions do not match well. With the semantic proximity $\mathcal{S}(\cdot)$ and the positional proximity $\mathcal{P}(\cdot)$, the full target similarities $\mathcal{T}(\cdot)$ for the weights in the differentiable similarity graph $\mathcal{G}$ can be formulated next.

Alignment to target similarity graph In order to restrict the similarities in our graph $\mathcal{G}$, we begin by introducing the function $\mathcal{T}(\cdot)$ to calculate the desired target

similarities between pairs of voxel embeddings. For a given edge (x, y) connecting voxel embeddings, the model is expected to generate a similarity value $\sigma(x, y)$ that aligns with the target similarity:

$$\mathcal{T}(x, y, \mathcal{R}(\cdot), p) = \alpha \cdot \mathcal{P}(x, y, \mathcal{R}(\cdot)) + (1 - \alpha) \cdot \mathcal{S}(x, y, p) \ . \tag{II.21}$$

The weight $\alpha \in [0, 1]$ enables us to balance the influence of the *receptive smoothness* and *semantic coherence constraints* in the optimization process. By utilizing the targets produced by $\mathcal{T}$, we can align the computed voxel embeddings of a given input volume to these desired similarity targets, thus promoting receptive smoothness and semantic coherence in our model training. We leverage the common contrastive similarity formulation for this alignment:

$$\mathcal{O}(e^x, e^y) = \frac{\exp(\sigma(x, y))}{\sum_{e^z \in \mathcal{N}} \exp(\sigma(x, z))} \ , \tag{II.22}$$

which encodes the voxel-embeddings and the similarities among them (*i.e.* the graph $\mathcal{G}$) that the current segmentation model produces. The loss function which can be optimized via back-propagation is formed by minimizing the cross-entropy between the computed targets $\mathcal{T}(\cdot)$ and the similarities $\mathcal{O}(\cdot)$:

$$L(\mathcal{Q}, \mathcal{N}) = - \sum_{e^x \in \mathcal{Q}, e^y \in \mathcal{N}} \mathcal{T}(x, y, \mathcal{R}(\cdot), p) \cdot \log(\mathcal{O}(e^x, e^y)) \ . \tag{II.23}$$

Lastly, we make the loss symmetrical by adopting the idea described in [111] to selectively back-propagate through either $\mathcal{Q}$ or $\mathcal{N}$. As a result, our proposed *Contrastive Constrained Regularization* loss function L_{Con2R} can be expressed as:

$$L_{Con2R}(\mathcal{Q}, \mathcal{N}) = \frac{1}{2} \cdot (L(\mathcal{Q}, \bar{\mathcal{N}}) + L(\mathcal{N}, \bar{\mathcal{Q}})) \ . \tag{II.24}$$

The notation of $\bar{\mathcal{Q}}$ and $\bar{\mathcal{N}}$ is intended to show that the respective voxel-embedding sets are detached from the computation graph, which means they are treated as

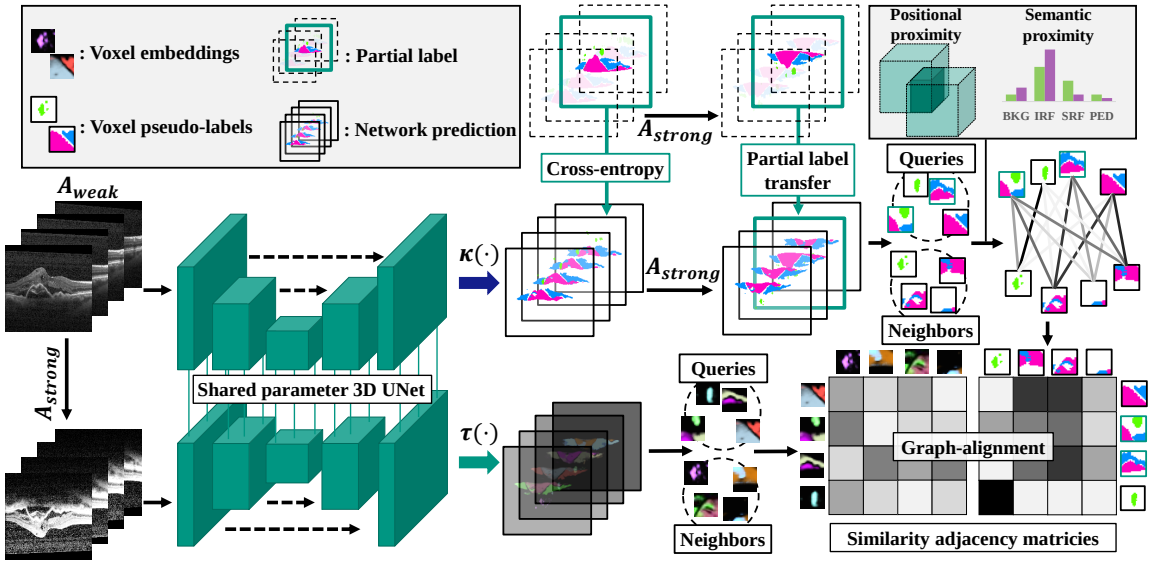


Figure 12: Our approach, referred to as *Con2R*, utilizes both weakly- and strongly augmented volumes to generate voxel-wise embeddings and construct a similarity graph with them. We aim to align this graph with a target similarity graph, which we compute based on positional- and semantic proximity constraints using network predictions as well as partial labels, if provided. This alignment procedure allows us to learn consistent 3D predictions solely using unlabeled and partially labeled data.

constants in back-propagation.

5.2.2 Graph-constrained semi-weak learning

Our loss formulation L_{Con2R} is the crucial part in our overall *Contrastive Constrained Regularization (Con2R)* training strategy, which we display in its entirety in Figure 12. With a naive semantic segmentation strategy, we train the 3D segmentation model by using weakly augmented volumes $\mathcal{A}_{weak}(v_i)$ as input and learn the model weights including weights in the output-head $\kappa(\cdot)$ by back-propagating a standard cross-entropy loss $L_{Entropy}$ (*cf.* Equation (II.6)) only for voxel-predictions that are associated to the few available partial annotations. In order to compute the similarity targets $\mathcal{T}(\cdot)$ in L_{Con2R}, we need semantic predictions p. These are gathered by simply processing the weakly augmented volume with the network and applying a

softmax function on top of the output of $\kappa(\cdot)$, yielding normalized class associations for each voxel. In case the input volume has associated partial labels, we further exchange the predictions at the labeled regions with the ground-truth and obtain the more precise semantic predictions p_i^*:

$$p_i^* = p_i \cdot (1 - a_i) + m_i \cdot a_i \ , \tag{II.25}$$

To make use of augmentation consistency in our training strategy, we compute the embedding graph $\mathcal{O}(\cdot)$ based on a strongly augmented version of the same volume $\mathcal{A}_{strong}(v_i)$, which necessitates a second forward pass through the network with the transformation output-head $\tau(\cdot)$ to produce voxel-wise embedding vectors. To be consistent and aligned, the geometric augmentations in $\mathcal{A}_{strong}(v_i)$ also have to be applied to the semantic predictions p_i^* as well. Then, with the correct alignment between voxel-embeddings and predictions, voxel locations can be sampled, *i.e.* the Query- and Neighborhood-sets can be set up, and the similarity targets can be computed with the location- and semantic prediction information. With this, all ingredients of L_{Con2R} are ready and the objective resolves to minimizing $L_{total} = L_{Entropy} + L_{Con2R}$.

5.3 Experiments and results

In this section, we begin by providing an overview of the datasets and experimental setup used to showcase the *Con2R* training strategy. To evaluate its effectiveness, we outline our protocol for training with partial annotations and discuss the methods we compare against. Lastly, we present quantitative and qualitative results, which prompt a discussion about the previously stated research questions in Figure 9.

5.3.1 Datasets

We assess the effectiveness of our method using two widely recognized volumetric datasets, the RETOUCH OCT dataset [8], which focuses on retinal fluid segmentation. This dataset consists of three classes: Intraretinal fluid (IRF), Subretinal

fluid (SRF), and Pigment Epithelium Detachment (PED). While multiple vendors of OCT devices are included in the dataset, we specifically concentrate on the Spectralis device, which has volumes composed of 49 b-scans. Additionally, we evaluate our approach on the task of brain tumor sub-region segmentation in magnetic resonance images. The data used for this evaluation is obtained from the medical decathlon data collection [7], which encompasses data from multiple BraTS challenges [238, 83, 82]. The tumor sub-regions we focus on are edema (EDM), enhancing tumor (EN), and non-enhancing tumor (NEN), within volumes of depth 155.

5.3.2 Evaluation protocol

Our focus lies on investigating scenarios where we have extremely few partial pixel-wise volume annotations, *i.e.* slice annotations, for which we outline the evaluation protocol next. The datasets encompassing volumes are split five times independently into train and test sets, whereas the training set of volumes is further divided into a train and validation portion of volumes, specifically for the RETOUCH dataset into 14 train, 5 validation and 5 test volumes and for the BraTS data 242 train, 121 validation and 121 test volumes. In coherence with the previously outlined procedure in Section 4.3.2, we report the mean and standard deviation of the performance of all models along these five-fold cross-validation results in order to respect the scarce data setting we are diving into. We evaluate the segmentation performance via the average mIoU measure, as defined in Equation (II.14), over the five splits. When setting up the train, validation and test sets, we make sure that all classes are covered in them. Further, for the train sets, we shuffle all train volume slices and enumerate them and then, subsequently build a sequence of scenarios, where only the first 3, 6, 12, or 24 slices are associated with pixel-wise annotations, with only a small condition, namely that all classes are present in the different scenarios. This naturally builds a sequence of a small partially labeled dataset which slowly grows bigger in terms of annotations. Thereby, we have scenarios where randomly selected slices, *i.e.* sub regions in the volumes, or, partial annotations are the basis

for training segmentation models, while all remaining slices are unlabeled, fulfilling our initial supervision design of Section 5.1.1. In terms of notation of this previous section, *e.g.* the scenario with 12 annotated slices, these annotations are distributed randomly among all volumes: $\sum_{i,d=1}^{N,D} a_i^d = 12$.

5.3.3 Implementation details

To train models in a semi-weakly supervised manner with partial annotations successfully, we notice that oversampling the partially labeled volumes to ensure, that in each iteration at least one labeled slice is present was important. Therefore, we constructed the batches of size two by always including a partially labeled volume (as also suggested in semi-supervised literature [231]) to not deviate to solutions mainly considering the predominant unlabeled volumes.

As neural volume segmentation architecture we chose the established 3D Unet [66] in a configuration with $64, 128, 256, 256$ channels in the encoder building blocks and the fitting reverse sequence in the decoder. The encoder building blocks are twice the sequence of: group normalization [239] with eight groups, convolution with kernel size three, padding one and ReLU activation function. To optimize the network weights we use SGD with a momentum term of 0.9, weight decay of 0.00001 and a learning rate of 0.01. We train a lower bound model which is only trained with cross-entropy loss on the partially labeled volumes [66] and does not consider the unlabeled volumes or unlabled portions of the partially labeled volumes. This lower bound model is initialized with Xavier initialization [229], while all other approaches are initialized with the weights of the respective lower bound model in the specific supervision scenario. The models are trained a total of 100 epochs and evaluated on the validation set every 10 epochs, with the best performing validation model being applied once to the test set after training to be evaluated.

We train the 3D Unets on volumes which are resized to $49 \times 160 \times 160$ for the RETOUCH dataset and to $155 \times 110 \times 110$ for BraTS. Afterwards, we crop sub volumes from the depth dimension to end up at $16 \times 160 \times 160$ and $32 \times 110 \times$

110 volumes for training on RETOUCH and BraTS, respectively. This process is necessary to train the networks with a batch size of two on 11GB NVIDIA RTX 2080 Ti GPUs.

The general weak data augmentations we employ is flipping the input volumes in longitudinal- and vertical directions in 50% of the cases. For approaches requiring strong augmentations including our Con2R approach (Section 5.2.2), we apply brightness and sharpness augmentations with randomly sampled magnitudes between $[0, 2]$ as well as an adapted CutOut [240] augmentation which cuts sub-volume chunks of $16 \times 16 \times 16$ from the input volume and sets them to zero. We found this augmentation configuration to work best for the datasets by thoroughly tuning baseline methods which we selected from semi-supervised 2D- as well as volume segmentation literature and adapted to our semi-weakly training scenario, and which we will outline in the following section.

5.3.4 Competing approaches

Standard 3D Unet [66]: The first baseline we train to set the lower bound is a 3D Unet where only the labeled regions in the partially annotated volumes influence the optimization through a cross-entropy loss as in [66]. This is the lower baseline that all other more sophisticated methods to train the 3D Unet architecture should be able to outperform.

Pseudo-label (PL) [99, 62]: To establish a first naive semi-supervised baseline for comparison, we adopt the widely used pseudo-labeling approach, a classical semi-supervised method and adapt it naively to volume segmentation. To set up pseudo-labels, we predict the semantic assignment (argmax of class predictions) for all voxels in all training volumes using the standard 3D Unet baseline and augment the volumes where we have partial annotation information using the methodology described in Equation (II.25). This equates to offline pseudo-labeling, which did not yield good results for our datasets and therefore, we refined this approach with the self-training normalization term of [62] which is described in Section B.1 of the Appendix.

Mean-Teacher (MT) [104]: We adopt the widely used Mean-Teacher framework, originally proposed for semi-supervised classification, and adapt it to handle volumetric inputs and dense predictions. The segmentation adaptation of this approach has been explored previously in [241, 33]. Our training procedure involves aligning the predictions of the student network with those of the teacher network, which are obtained by forwarding differently augmented volumes through the networks. To ensure pixel alignment between the student and teacher outputs for the consistency loss, we apply reverse geometric augmentations on the teacher predictions to match the output of the student. This is consistent to the procedure previously introduced in Section 4.2.3. Determining the exponential-moving average decay factor empirically led us to set $\alpha = 0.5$ for the best results, consistent with our previous investigation in Section 4.3.5, Table 1 for the 2D segmentation case.

Uncertainty-aware Mean-Teacher (UA MT) [125]: In this variant of the Mean-Teacher approach, uncertainty estimation using Monte-Carlo dropout [242] was introduced to selectively apply the consistency loss between student and teacher. This is done by computing the voxel-wise uncertainty of the model and threshold it, to only apply the consistency loss to volume regions below the threshold. Our experimentation, which is available in Section B.2, determined a threshold value of 0.5 and performing 8 forward passes for the Monte-Carlo dropout to yield the best results.

FixMatch [102]: FixMatch is a highly effective method primarily developed for 2D classification tasks. It combines pseudo-labeling with consistency regularization by incorporating both weak and strong augmentations as described earlier. As we apply this approach to segmentation, we take into account the alignment of predictions between the strongly augmented branch and the weakly augmented branch and employ a similar strategy to what we describe for the Mean-Teacher. In our experiments, we found that using a confidence threshold of 0.5 on the network predictions from the weakly augmented branch yielded favorable segmentation results.

Contrastive Constrained Regularization (Con2R): Our own *Con2R* method is trained by sampling Query- and Neighborhood sets of size $|\mathcal{Q}| = |\mathcal{N}| = 1,728$

$\mathcal{R}$	validation mIoU
$16 \times 16 \times 16$	$49.1 \pm 4.7\%$
$32 \times 32 \times 32$	$47.1 \pm 2.7\%$
$64 \times 64 \times 64$	$46.3 \pm 2.3\%$
$160 \times 160 \times 160$	$46.5 \pm 6.1\%$

Table 5: Effect of receptive volume size $\mathcal{R}$ on the mean IoU

$\lvert\mathcal{Q}\rvert, \lvert\mathcal{N}\rvert$	validation mIoU
216	$46.9 \pm 4.5\%$
512	$46.9 \pm 3.4\%$
$1,000$	$47.8 \pm 5.0\%$
$1,728$	$49.1 \pm 4.7\%$

Table 6: Effect of number of vertices in graph $\mathcal{G}$ on the mean IoU

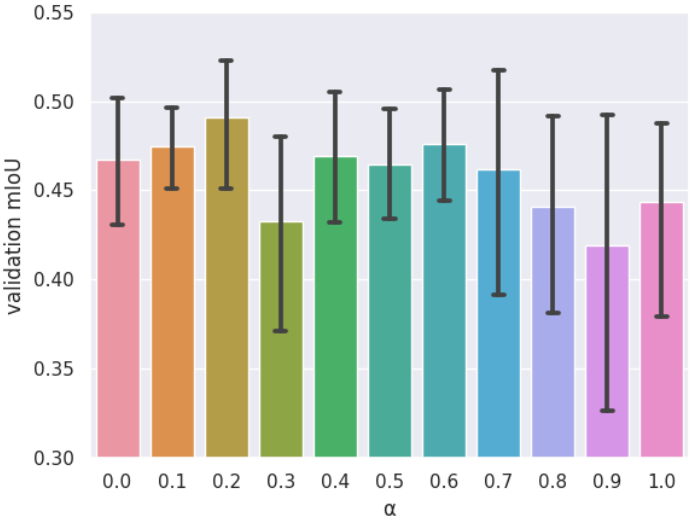


Table 7: Validation performance of *Con2R* when tuning α, IoU reported along five validation splits with mean and standard deviation displayed

for RETOUCH and a size of $3,375$ for BraTS. This relates to the maximal number of voxel-embeddings possible to be sampled within our computational budget. The interpolation between the positional- and semantic constraints for setting up the target graph is controlled by $\alpha = 0.2$. The receptive volume size $\mathcal{R}$ is $16 \times 16 \times 16$ and $32 \times 32 \times 32$ for RETOUCH and BraTS, respectively. For the embedding graph, we compute $e_{dim} = 64$ dimensional voxel-embeddings and smooth the target similarity graph with $\varepsilon = 10^{-7}$.

5.3.5 Hyper-parameter sensitivity studies

We study the sensitivity of our *Con2R* method to its hyper-parameters by experimenting in the scenario of having 24 annotated slices among the training volumes on the RETOUCH data. In Table 5, we present the impact of selecting the receptive field size in $\mathcal{R}$ on the accuracy of volume segmentation. The highest accuracy is achieved when using a receptive field size of $16 \times 16 \times 16$, which corresponds to the maximum depth of the input volume crops. Thus, as we train with volumes that are

	Partial Volume Supervision				
Method	3	6	12	24	Full Access
3D Unet [66]	12.0 ± 5.6	18.1 ± 11.5	31.1 ± 12.4	43.8 ± 2.5	54.9 ± 0.9
PL [99, 62]	13.0 ± 6.3	20.6 ± 13.4	30.9 ± 11.5	45.7 ± 2.2	55.4 ± 1.5
MT [241, 104]	12.0 ± 6.6	20.2 ± 12.4	34.4 ± 11.4	45.3 ± 3.1	53.4 ± 1.9
FixMatch [102]	10.4 ± 5.7	18.7 ± 10.6	34.7 ± 6.8	46.2 ± 3.8	54.4 ± 3.3
UA MT [125]	13.0 ± 6.7	20.0 ± 11.9	36.5 ± 9.2	45.7 ± 1.9	$\mathbf{56.3 \pm 1.7}$
Con2R (Ours)	$\mathbf{14.8 \pm 8.7}$	$\mathbf{22.5 \pm 10.0}$	$\mathbf{38.6 \pm 7.5}$	$\mathbf{48.2 \pm 3.1}$	54.6 ± 1.2

Table 8: RETOUCH results in mIoU for semi-weakly supervised learning, number of annotated b-scans successively increased from 3 to 24 and full access as upper limit

32 slices deep for the BraTS task, we adjust $\mathcal{R}$ to $32 \times 32 \times 32$. It is worth noting that larger receptive field sizes result in a degradation of performance, and we hypothesize that the shape and size of objects in the dataset being segmented also play a crucial role in determining an appropriate choice. In Table 6, we explore the effect when varying the number of sampled vertices in the Query- and Neighborhood-set from the volume-graph. Increasing this number consistently enhances the effectiveness of *Con2R*. To accommodate the available GPU capacity, we simply set this hyper-parameter to the maximum values of $1,728$ for the RETOUCH dataset and to $3,375$ for BraTS. Lastly, we investigate the influence of the weight α on the relationship between the positional- and semantic constraints (Table 7). We observe that semantic constraints alone ($\alpha = 0.0$) yield favorable outcomes, while surprisingly, positional constraints alone ($\alpha = 1.0$) also lead to solid results. However, the optimal performance is achieved with $\alpha = 0.2$, putting more weight on the semantic constraint part of Equation (II.21). Ablations of baselines can be found in Appendix Section B.

5.3.6 Quantitative results

Equipped with methods which have good hyper-parameters for learning with partially annotated volumes and experimental settings to test their efficacy in low supervision scenarios, we investigate in Table 8 how they perform with 3, 6, 12, 24

annotations on the RETOUCH retinal fluid segmentation dataset. The performance ceiling is given by the *Full Access* scenario situated furthest to the right. There, the mIoU values range between 53.4% and 56.3% with the lowest performance by the Mean-Teacher baseline, indicating that the consistency loss does not add important information in the training when fully annotated volumes are present, while the highest performance is achieved by the Uncertainty-aware Mean-Teacher which alterates the 3D Unet architecture by the integration of dropout layers which have a positive effect in this scenario. What we are most interested in, is the behavior of the segmentation models when they are supplied with very few annotations, as this brings us closer to answers for the research questions of this chapter (Figure 9). Thus, we can look into the lower baseline, *i.e.* training a 3D Unet model with only partial volume annotations naively [66]. The initial segmentation performance, as expected when training a model with as few as 3 slice annotations, remains very low at 12.0% mIoU. This can be broken down to 4.0% mIoU per slice annotation. When 3 more annotations are added, the mIoU increases by 6.1% absolute mIoU, which boils down to 2.0% per added annotation from the 3 to the 6 annotation scenario. Adding further 6 and 12 annotations (scenarios 12 and 24), the added value brought about by each annotation diminishes to 2.1% or 1.1% mIoU, while of course the absolute performance increases. A more annotation-efficient segmentation algorithm should produce a high absolute performance with few annotations. Framed differently, the added value of each slice annotation, should be higher for early annotations, *i.e.* the first few annotations should each have a bigger impact on the segmentation performance. For our Con2R method, this is the case, as for the four scenarios, it successively adds 4.9% per annotation in the 3 scenario, still 2.6% for the next 3 annotations in the 6 scenario while the next 6 more examples increase the performance each by 2.7%. In the scenario where Con2R is trained using 24 annotations, the individual contribution of the added annotations 12 annotations as compared to the previous scenario gets smaller to 0.8% per annotation, while the absolute performance of Con2R is still the best among the compared methods with

Method	Partial Volume Supervision											
	3			6			12			24		
	IRF	SRF	PED	IRF	SRF	PED	IRF	SRF	PED	IRF	SRF	PED
3D Unet [66]	21.1	11.5	3.3	21.0	24.6	8.6	23.4	49.9	20.0	30.5	73.0	27.8
PL [99, 62]	**22.4**	13.0	3.7	**23.5**	27.3	11.0	24.2	52.9	15.6	32.3	73.9	30.8
MT [241, 104]	18.4	12.9	4.8	20.7	29.4	10.5	24.5	59.7	19.0	30.9	76.6	28.3
FixMatch [102]	16.5	13.1	1.4	21.4	27.7	7.0	20.0	64.9	19.2	**33.4**	76.8	28.3
UA MT [125]	22.3	12.9	3.7	21.1	29.6	9.4	27.2	61.1	21.2	31.9	75.9	29.3
Con2R (Ours)	20.2	**16.4**	**7.8**	22.1	**31.8**	**13.6**	**27.3**	**65.2**	**23.3**	31.6	**79.1**	**34.0**

Table 9: RETOUCH class-wise results in mIoU for semi-weakly supervised learning, number of annotated b-scans successively increased from 3 to 24

48.2% mIoU due to the steeper performance gains early on. A broad view of the results of the adapted semi-supervised methods shows, that semi-supervision is with exception of two low supervision scenarios always beneficial. The two scenarios which fall below the lower baseline are FixMatch trained with 3 annotations, which might well be due to the severity of the training scenario and the Pseudo-label method trained with 12 annotation masks, which shows its general instability in training, as also discussed in Section B.1. In terms of absolute performance, Con2R outperforms the other baselines by +1.8%, +1.9%, +2.1% and +2.0% average mean IoU in the scenarios using 3 through 24 annotations. Table 9 gives a deeper view into these performance differences and where they come from. There, the individual disease classes are listed with their respective IoU. We can directly see, that our Con2R model performs considerably better than all competing approaches on the **Subretinal Fluid** and **Pigment Epithelial Detachments**. Qualitatively, these classes cover larger, connected areas in the volumes while **Intraretinal Fluid** often occupies small spots and tends towards more sprinkled regions. This can be explained with the hypothesis that larger classes profit more from the assumption that close voxels should resemble similar embeddings (positional proximity assumption), as this is more often true if voxels are surrounded by semantically similar voxels, *i.e.* large semantic areas. Yet, compared to the baseline 3D Unet, Con2R's way of training

the models and enforcing consistency among differently augmented input volumes still helps in segmenting **Intraretinal Fluid** better for the scenarios 6, 12 and 24, merely falling short for this class when only 3 annotations are available during training. The efficacy of our method regarding the **Subretinal Fluid** class can bee seen in the 24 annotation case, there Con2R segments **Subretinal Fluid** with an IoU of 79.1% which is quite close to the best fully supervised result of 84.4%, reached by Uncertainty-aware Mean-Teacher, which had access to the full 686 annotated slices.

A second task which we investigate our algorithms on is brain tumor sub-region segmentation, *i.e.* training models on the BraTS data. On this data, we experiment with a scarce annotation scenario using only 24 annotated slices among the volumes as well as quantify the Full Access upper bound in Table 10. In this setting, when training with 24 annotated slices, there is an even more extreme imbalance between annotated regions and unlabeled regions, namely 37,486 of the slices in the volumes are unlabeled. Still, all semi-supervised segmentation methods outperform the 3D Unet baseline with Con2R outperforming it with an absolute +4.4%. Compared to the best semi-supervised method, which is FixMatch, our method adds another +0.8% mIoU. The performance of Con2R with a fraction of 0.06% annotations compared to the Full Access scenario, already amounts to 79.1% of the Full Access performance. This again highlights that the lion's share

Method	Partial Volume Supervision				
	24				Full Access
	EDM	EN	NEN	mean	mean
3D Unet [66]	48.7	19.6	48.1	38.8 ± 3.4	51.7 ± 7.0
PL [99, 62]	49.1	21.3	50.5	40.3 ± 2.5	52.2 ± 8.4
MT [241, 104]	49.1	21.7	45.0	38.6 ± 4.5	53.7 ± 5.7
FixMatch [102]	50.1	**24.2**	53.1	42.4 ± 4.9	51.0 ± 6.5
UA MT [125]	49.2	22.6	51.3	41.1 ± 3.5	52.6 ± 6.0
Con2R (Ours)	**51.8**	23.9	**53.9**	$\mathbf{43.2 \pm 3.5}$	$\mathbf{54.6 \pm 7.7}$

Table 10: BraTS class-wise results in mIoU for semi-weakly supervised learning, number of annotated b-scans is set to 24 and full access is shown as upper limit

of annotation effort goes into lifting the performance ceiling higher and the growing impediment of achieving it with human labor, giving reason to developing better segmentation models in general as well as making them more annotation-efficient.

5.3.7 Qualitative results

Next, we gather some qualitative insights of the trained segmentation models, starting with the results on the optical coherence dataset RETOUCH in Figure 13. There, we see, that training with partially annotated volumes using merely 3 annotated slices in total, leads to very poor segmentation results for all approaches. The only fluid type which starts to be recognized in the correct location is (the blue) **Subretinal Fluid**. Adding 3 more slice annotations only helps the Con2R model in starting to grasp the **Pigment Epithelial Detachments** in pink, while the remaining methods incorrectly predict the green **Intraretinal Fluid** class, but still, results are quite poor overall. The baselines FixMatch and Uncertainty-aware Mean-Teacher correctly pick up the spatial relations between **Pigment Epithelial Detachments** and **Subretinal Fluids** when supervised with 12 partial annotations, just like Con2R. Yet, where Con2R comes out on top is in the 24 scenario, where we see how the data-driven constraints which we integrated into the training procedure really help in forming consistent segmentations. Here, we see that all other approaches produce speckled predictions, where **Subretinal Fluid** predictions leak into the **Pigment Epithelial Detachement** prediction, while due to the positional coherence constraint, our method produces smooth semantic predictions without such semantic region leakage.

For qualitative results on the BraTS dataset, we present segmentations for the semi-weakly supervised scenario where only 24 slices are annotated, in Figure 14. As also indicated by the previous quantitative results, 3D Unet and the plain Mean-Teacher methods either over-segment or severely under-segment the sub-tumor regions. The other competing methods over-segment the green **edema** class notably while training our Con2R model could capture details such as the small split at the bottom of the

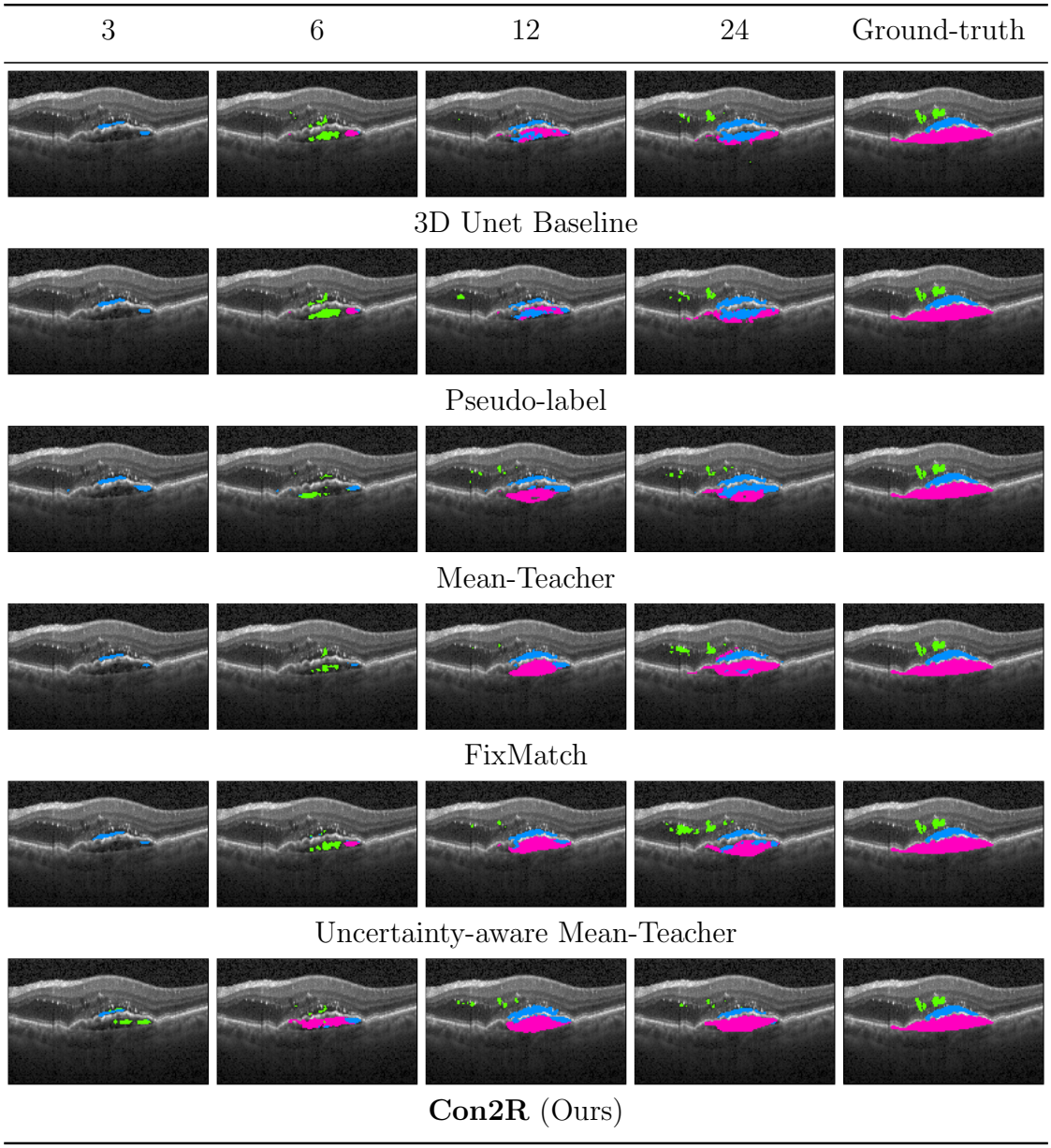


Figure 13: Segmentation progression when increasing the number of annotations from 3 to 24 in semi-weak retinal fluid segmentation, results for **IRF**, **SRF** and **PED** overlayed with input OCT scan. Right column: ground-truth.

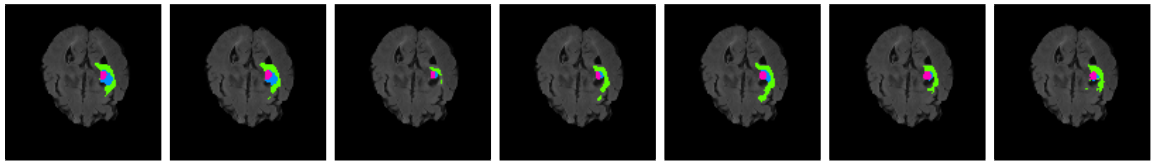


Figure 14: Segmentation results with 24 annotations in semi-weak brain tumor sub-region segmentation, results overlayed with first input channel of MRI scan

edema area.

A qualitative side-effect of Con2R is the possibility to use the learned embeddings of the trained transformation head $\tau(\cdot)$ to propagate a semantically labeled slice through the volume, as shown in Figure 15. There we use a single slice annotation from a previously unseen volume from the test set to propagate the semantic information through the whole volume. The label propagation can be achieved by searching, for each unlabeled voxel, the most similar voxel-embedding among the voxels of the annotated slice. Then the class of the most similar annotated voxel is assigned. The similarity computation is done via the cosine similarity. In the figure, the marked ground-truth slice is used to propagate its semantic information through the whole volume, displayed in the second row. For adjacent slices the propagated segmentation is reasonable, though it deteriorates the further away the unlabeled voxels lie in the volume as evident by the faulty segmentation of slices to the far right. In the third row the predictions of the standard output-head is shown, which for this example fails to segment the pink **Pigment Epithelial Detachment** class. On the other hand, the simple propagation strategy based on a query slice can only propagate class information present in the annotated slice. An additional visualization concerning the voxel-embeddings of a trained model can be found in Section B.4.

5.4 Discussion

When faced with segmentation tasks for volumetric data, a big challenge in profiting form the additional spatial dimension to learn diverse 3D patterns weights heavy on

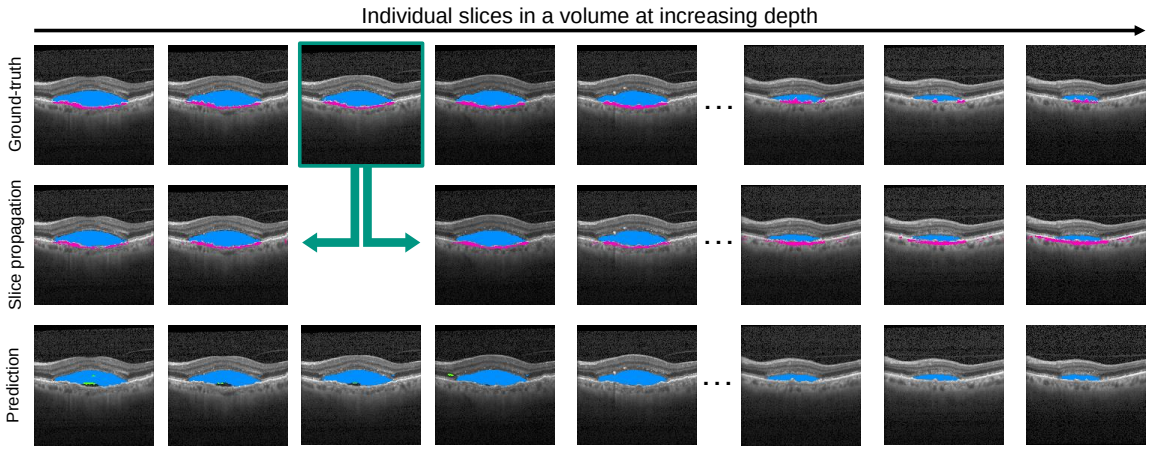


Figure 15: Qualitative segmentation for an OCT volume using a Con2R model trained with 24 slice annotations. The first row shows the ground-truth annotation of the volume, the second row is the resulting segmentation when propagating the marked ground-truth slice annotation via voxel-embbeding similarities, the third row shows standard predictions by the same trained Con2R model.

the annotation process: the additional dimension increases the annotation effort for each training example. Especially in expert-driven domains, where time to annotate is scarce, annotating whole volumes to train segmentation models becomes a big hurdle. What happens if we supply volumetric segmentation models only with sparsely populated, partial volume annotations? In the above experimental investigation, we gave insight into models trained in such a scenario and explored how they behave with extremely few slice annotations. We uncover that naively training volume segmentation models is outperformed by semi-supervised algorithms, which we adapt to accept partially annotated volumes aside of unlabeled volumes. Yet, semi-supervised volume segmentation algorithms leave behind some uncollected rewards, which we found out by designing a proxy-task that specifically respects the nature of partial annotations and enforces data-driven constraints that are rooted directly in the segmentation task. With advantageous effects on the segmentation performance when training with this proxy-task, we enable experts to, rather than providing dense annotations, cover a wider diversity of volumes with only sparse annotations. Further, models can already be trained with very few slice annotations and therefore, the

whole pipeline of annotation and segmentation model training can be designed in an iterative fashion. What our method can further contribute to such a procedure is the possibility to use the voxel-embeddings from a model trained on small amounts of annotations as tool to propagate the slice annotation information to adjacent frames of the volume data, potentially easing the annotation of volumes. The flexible training strategy in our Con2R method enables learning from more convenient partial labels and unlabeled data, easing the process for expert annotators. At the same time, the design to alter the segmentation architecture to include a second voxel-embedding output branch and our loss function to ensure positional- and semantic coherence among the voxel-embeddings, offers potential for a future redesign of the annotation processes towards an interactive- and expert-centric design.

Contribution 1:
We investigated volume segmentation scenarios where algorithms are supplied with partial volume annotations as well as completely unlabeled volumes. In this setting, we evaluated the performance of current semi-supervised algorithms and how their behavior changes with more and more annotated regions distributed over the training set volumes.

Contribution 2:
We proposed the Contrastive Constrained Regularization training strategy, where we designed a positional proximity- as well as a semantic coherence constraint with the aim to overcome the mismatch between dense volumetric predictions and sparse, partial annotations, which are supplied in training. With this we achieve the best semantic segmentation performance on two medical datasets in optical coherence- and magnetic resonance imaging and show that our Con2R method profits most from additional slice annotations in extremely scarce supervision scenarios.

In the next chapter, we consolidate the insights we have gathered in chapter 4 and chapter 5 and design an algorithms which is able to profit from as diverse annotations

as pixel-wise annotations, bounding boxes, single points, image-level labels as well as entirely unlabeled images. There we build on the insights we gathered in previous chapters, *e.g.* the idea of pseudo-label filtering, the siamese architecture design as well as the view through trained embeddings on the segmentation task. We side-line the introduction of a novel method to learn from diverse, heterogeneous semantic annotations by more systematic means to analyze semi-weakly supervised algorithms.

6 Unified learning with diverse annotation types

The limiting factor for bringing semantic segmentation solutions to expert-driven domains is the process of acquiring annotations. In order to ease annotation and center it more around expert-annotators, we aim at, from the algorithmic side, accepting diverse annotation types, making it possible for experts to spend their time more efficiently on annotations of different granularity and thereby of a varying expenditure of time. The question we consider in this chapter is how segmentation models can be trained with such diverse annotations, which we answer by designing the *Decoupled Semantic Prototypes (DSP)* method. DSP is a semi-weakly supervised contrastive loss term which unifies learning from as diverse signals as masks, bounding boxes, points, image-level labels as well as unlabeled data. By also proposing the notion of an *Annotation Compression Ratio*, we are able to quantify what mixtures of annotation types bring the highest performance, and uncover implications for time-saving in the annotation process. Our segmentation solution DSP and analysis of its performance in varying supervision scenarios has direct effects on how the expert's time can be used more effectively for bringing semantic segmentation into expert domains.

This section is based on a publication in *CVPR 2023* [35], experiments were done on the HoreKa supercomputer funded by the Ministry of Science, Research and the Arts Baden-Württemberg and by the Federal Ministry of Education and Research.

6.1 Introduction

In the previous chapters training strategies for segmentation algorithms with varied, but specific annotation types were presented to ease the annotation process for experts and redesign it from an algorithmic perspective. Such semi-weakly annotation combination included pixel-wise masks and unlabeled images, partially labeled volumes plus unlabeled volumes and masks combined with image-level labels. As

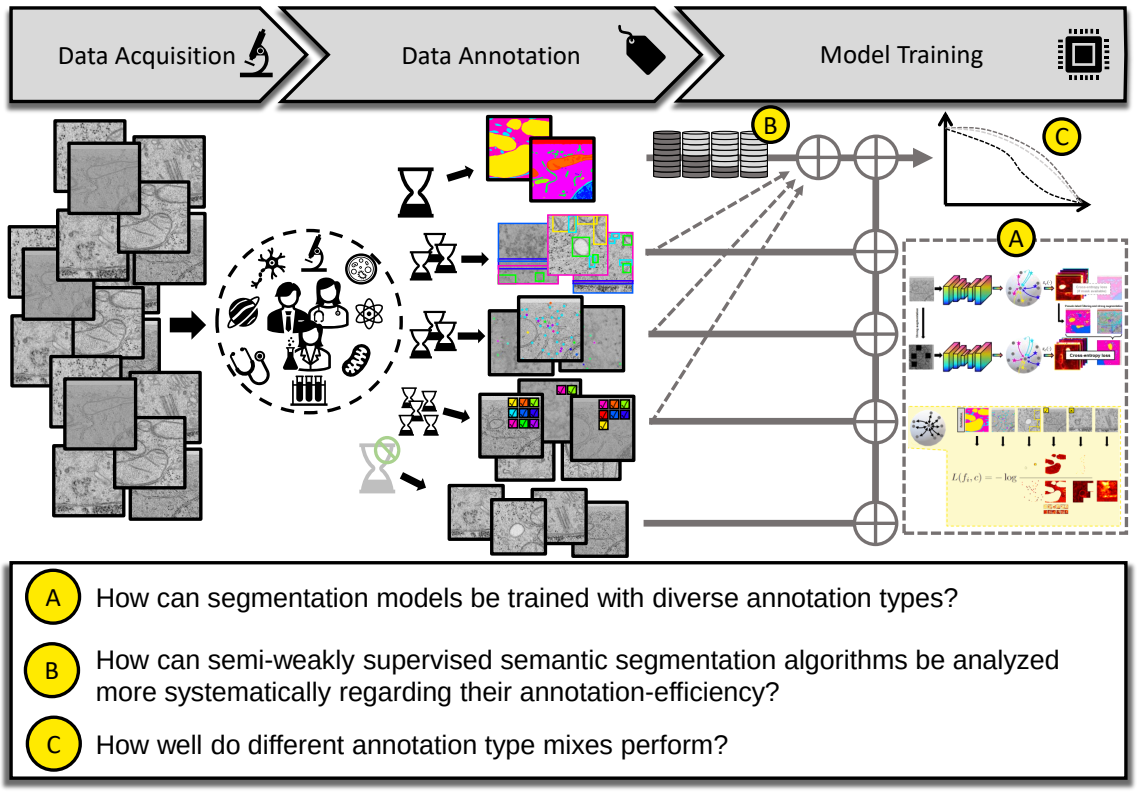


Figure 16: Overview of the main research questions in this chapter, they will be explored on an electron microscopy dataset [9] where biologists are needed in the annotation process. We further outline the *Decoupled Semantic Prototype* method which can be trained with diverse annotation types making it flexible and putting expert annotators into the center of the pipeline.

we outlined in the respective chapters, this already enables for more diverse annotations in the segmentation network training procedure which frees experts to more flexibly spend their time for annotations. Yet, an open question remains: Can a model be trained with an even more diverse set of annotation types to free the experts even more from annotation restrictions? In this chapter we investigate this fundamental question and design a segmentation method which is able to make use of pixel-wise masks, bounding boxes, point annotations, image-level labels as well as unlabeled images at once. Thereby, our training procedure is able to accept the

most prominent semi- and weak supervision signals from literature at once. With this algorithm, we can for the first time shed light on how well a model performs with such diverse annotation types mixed together as opposed to just pairs of annotations in isolation. As in previous chapters, we are especially interested in the performance of segmentation algorithms, when they are trained with only very few pixel-wise annotations, and how the performance evolves when successively adding more. To this extent, we explore a more systematic procedure to investigate this by introducing the notion of an *Annotation Compression Ratio*, with which a model is trained. With this more systematic approach to measure an individual segmentation algorithm's performance by exponentially reducing the amount of pixel-wise annotations used to train a model, we can gather interesting insights. Further, by using different annotation type combinations, we can investigate which annotation type mixes perform best, and whether heterogeneous annotation types in a training set are a disadvantage or might even be a better, more economical choice. With the design of the new segmentation algorithm as well as the segmentation results for different annotation mixes implications on the annotation process for expert-driven segmentation datasets can be drawn. The main research targets of this chapter are summarized in Figure 16.

6.1.1 Problem statement

To train segmentation models, we define a training dataset $\mathcal{D} = \{x_1, \ldots, x_n | x_\ell \in \mathbb{R}^{c_{dim} \times H \times W}\}$ comprised of images x_ℓ with dimensions $\mathbb{R}^{c_{dim} \times H \times W}$. Here, c_{dim} represents the number of color or intensity channels, while H and W denote the height and width of the images, respectively. Our training approach is designed to accommodate a variety of annotation types, allowing for a broad semi-weakly supervised segmentation scenario. In this scenario, which is depicted in Figure 17, images x_ℓ can be either unlabeled $\mathcal{U}$, accompanied by pixel-wise masks $\mathcal{M}$, weakly annotated with bounding boxes $\mathcal{B}$, point annotations $\mathcal{P}$, or image-level labels $\mathcal{I}$. Annotation types are defined as previously described in Section 4.1.2, except for bounding boxes

and point annotations which we refer to as either two coordinates $(u_1, v_1), (u_2, v_2) \in [0, \ldots, H-1] \times [0, \ldots, W-1]$ on the image defining a box accompanied by a class association $\in [0, 1]^C$ or a single coordinate, *i.e.* point, associated with a class label, respectively. It is worth noting that images annotated with pixel-wise masks provide access to weaker annotation types such as point, box and image-level labels, as they can be derived from them, while box- and point annotated images inherently contain information about the image-level label of the image.

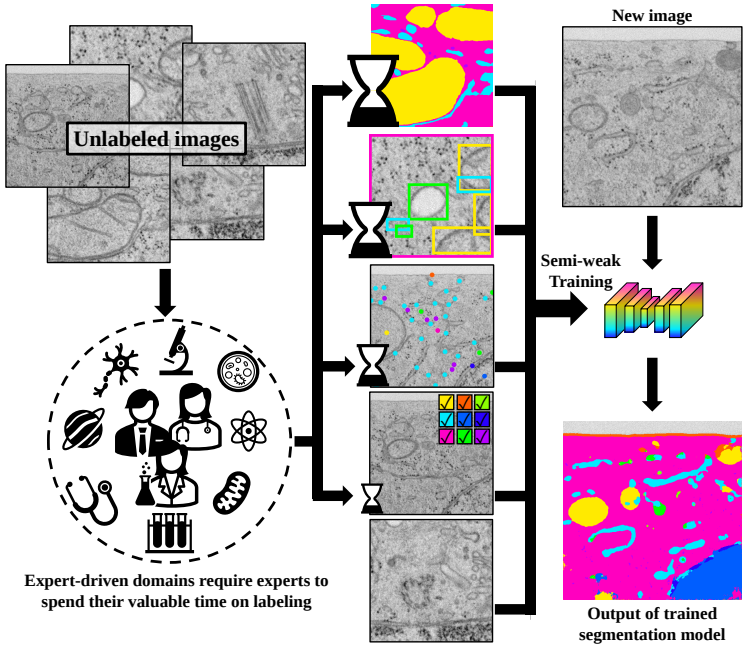


Figure 17: Algorithms which can be trained with as diverse annotation types as masks, boxes, points, image-level labels and unlabeled images can help in facilitating the annotation time of experts more conveniently.

6.1.2 Preliminaries

A key idea of our method is to extend our previous idea of chapter 5, *i.e.* using pixel-wise embeddings, and enforce a contrastive loss function which can work with diverse semantic cues based on different annotation modalities and thereby enforce dependencies on a pixel level. Hence, it is necessary to assign an embedding vector to each pixel in the image. In contrast to conventional segmentation networks that employ a 1×1 convolution and pixel-wise cross-entropy loss for training, we propose a network architecture ε that generates embeddings $F \in \mathbb{R}^{D \times H \cdot W}$ consisting of embedding vectors $f_i \in \mathbb{R}^D$ corresponding to each pixel i. Such an architecture can

be obtained by simply omitting the last 1×1 convolution from any given network, making it applicable to previous segmentation architectures. Next, we introduce semantic prototypes, which can be used to map the individual pixel-embeddings f_i to the semantic classes again.

6.2 Decoupled semantic prototypical networks

While operating on pixel-wise embeddings will enable us to design a contrastive loss function in a D dimensional embedding space to profit from different annotation types, first representations of the semantic classes in this space are needed and a way to associate individual pixel-embeddings to them. Therefore, we make use of semantic prototypes $p_c^j \in \mathbb{R}^D$, which are learnable parameters in the form of vectors. Here, c indicates the class which is represented by the prototype, while j indexes the specific prototype in a set of prototypes P_c for this class. To obtain a class prediction for a pixel-embedding f, we compute the cosine similarity of it to all prototypes p_c^j:

$$\sigma(f, p_c^j) = \frac{f^\top p_c^j}{||f|| \cdot ||p_c^j||} \ . \tag{II.26}$$

This formula yields the similarity between a single prototype of a class c to a single pixel-embedding f. To end up at a class-score for f the mean similarity to all prototypes of class c, *i.e.* the set P_c, has to be computed:

$$s_c(f, P_c) = \frac{1}{|P_c|} \cdot \sum_{j \in P_c} \sigma(f, p_c^j) \ . \tag{II.27}$$

With this aggregation, the score s_c quantifies how similar a pixel-embedding is to all prototypes representing class c. To normalize these scores, we apply a temperature scaled softmax function:

$$\bar{s_c}(f, P_c) = \frac{\exp\left(s_c(f, P_c)\right)/\tau}{\sum_{i=1}^{C} \exp\left(s_c(f, P_i)/\tau\right)} \ , \tag{II.28}$$

where τ is the temperature parameter. The scores $\bar{s_c}$ can be interpreted as class predictions similar to classical predictions produced by a conventional 1×1 convolutional output-head. Thus, $\bar{s_c}$ can be used in conjunction with a cross-entropy loss when a pixel-wise annotation is available to train the segmentation network. In order to infer the segmentation of an image at inference time, pixel-embeddings need to be computed using the segmentation network ε to obtain $f \in F$ and associated to the classes by computing $\bar{s_c}$ for all C classes. Then, the class c with the highest score for each pixel-embedding is set as the predicted class: $\arg\max_c \{\bar{s_c}(f, P_c)\}_{c=1}^{C}$.

A note on semantic prototypes All prototypes in P are learned end-to-end via loss functions and back-propagation. Although the mathematical formulation above, at first glance, looks quite different from the standard 1×1 convolutional output-head it is quite similar. As all prototypes can be put into a single matrix of shape $D \times |P_c| \cdot C$ and via a matrix-multiplication for the similarity computation with the pixel-embeddings F of shape $D \times H \cdot W$ it would be equivalent to applying $|P_c| \cdot C$ 1×1 convolutions to each pixel-embedding. The main difference to the standard segmentation procedure lies in that we have $|P_c|$ outputs per class and the prototypes are normalized, which would equate to the weights within the $|P_c| \cdot C$ 1×1 convolutions being normalized. This connection relates our semantic prototypes to weight normalization as proposed by Salimans *et al.* [243] with the exception, that there arbitrary magnitudes $\neq 1$ of the weight vectors are allowed, while we stay at unit length for each prototype and we further normalize the input pixel-embeddings.

In a more descriptive way, our prototypes can be thought of as implicitly finding semantic cluster centers which differentiates them from the proposition of Zhou *et al.* [244] where online clustering is applied to explicitly obtain prototypes for segmentation. There is also a coarse relation to class queries as used in some segmentation transformer architectures [245]. With our altered view on segmentation, it can also be thought of as, instead of predicting one-hot categorical vectors directly, our model learns in a data driven way D-dimensional class representations that encompass the

same embedding space as representations of pixels. On top, due to the usage of multiple semantic prototypes per class this modelling allows for classes to be represented by a multi-modal distribution of representation vectors in this embedding space.

6.2.1 Decoupled semantic prototypes

While the semantic prototypes can be trained end-to-end via enforcing a cross-entropy loss on the scores $\bar{s}_c(f, P_c)$, this requires pixel-wise annotations for pixel-embeddings f which we might only have access to very few. What the design choice of Sec. 6.2, *i.e.*, working on an embedding space and on similarities between pixel-embeddings and class-representative prototypes enables is a well-directed manipulation of that embedding space. Such a directed manipulation can be achieved via the tool of contrastive learning, where associations between certain pairs of representations are enforced and their cosine similarity increased, while other associations get weakened by penalizing a high cosine similarity among them. In constrastive literature, associations that should be strengthened are termed positive pairs while associations that should be weakened are referred to as negative pairs [246, 247, 113, 114]. Generally, contrastive loss terms are built by using different data augmentations on a single image and process it via a network to obtain two high dimensional feature representations of the same image z_i and $\hat{z}_i$. The idea is that because the representations stem from the same image, merely altered with an augmentation that largely preserves the semantic content, they can be deemed as positive pairs while all associations between z_i and a representation z_j obtained from another image is seen as negative. In terms of a loss function, we adapt the notation of [114] and write:

$$-\log \frac{exp\left(\sigma(z_i, \hat{z}_i)/\tau\right)}{Z_i} \quad , \tag{II.29}$$

with the positive association in the numerator and Z_i denoting the negative pairs with respect to z_i. In a batch of size B the standard contrastive loss as formulated

e.g. in [113] can be noted down as:

$$Z_i = \exp\left(\sigma(z_i, \hat{z}_j)/\tau\right) + \sum_{j=1, j\neq i}^{B} \exp\left(\sigma(z_i, \hat{z}_j)/\tau\right) + \exp\left(\sigma(z_i, z_j)/\tau\right). \tag{II.30}$$

Here, the denominator, *i.e.* the set of negative pairs is made up of all pairs between the positive representation z_i and all remaining vectors, except for the pair with itself (z_i, z_i). Thus, this contrastive formulation can be seen as a classification task with the individual instance representations as the targets and only one instance, $\hat{z}_i$, has the label 1 while all other instances, in the denominator, have the label 0. Yeh *et al.* propsed to completely decouple the denominator from the pair which encodes the same instance in their Decoupled Contrastive Loss (DCL) formulation:

$$Z_i = \sum_{j=1, j\neq i}^{B} \exp\left(\sigma(z_i, \hat{z}_j)/\tau\right) + \exp\left(\sigma(z_i, z_j)/\tau\right). \tag{II.31}$$

Here, both pairs (z_i, z_i) as well as the augmented pair $(z_i, \hat{z}_i)$ are omitted from the denominator. This DCL adjustment led to stronger self-supervised performance for small batch sizes. We modify the idea of decoupling the numerator and the denominator in that we intent to decouple them not based on a single instance which has been augmented twice, but we completely decouple the denominator from the class association of the pair in the numerator. But first, we outline which representations are contrastively paired and input in the cosine similarity computations $\sigma(\cdot)$. In Section 6.2, we modified arbitrary segmentation architectures by omitting the penultimate 1×1 convolution output-head, operating directly on pixel-embeddings and adding semantic prototype vectors which are associated to the classes. In our variant of the contrastive loss function, we insert associations between pixel-embeddings f_i and the prototypes p_j, or more precisely the association of a pixel-emebdding to the set of semantic prototypes for a specific class $s_c(f_i, P_c)$. Adapting Equation (II.29)

results in:

$$L(f_i, c) = -\log \frac{\exp\left(s_c(f_i, P_c)/\tau\right)}{Z_{i,c}} \quad . \tag{II.32}$$

The normalization factor $Z_{i,c}$ now has to include all $B \cdot H \cdot W$ pixels in all images of the batch, which are associated to all prototypes of the C classes:

$$Z_{i,c} = \sum_{j=1}^{B \cdot H \cdot W} \sum_{k=1, j \neq i \wedge k \neq c}^{C} \exp\left(s_k(f_j, P_k)/\tau\right) \quad . \tag{II.33}$$

With the condition $j \neq i$, we follow the design of the DCL formulation, in order to decouple the positive instance, here $s_c(f_i, P_c)$, in the numerator from the denominator. Yet, leaving the condition as is would omit all associations between the positive pixel-embedding f_i and classes other than c, therefore we extend the condition to $j \neq i \wedge k \neq c$ which explicitly rules out $\exp\left(s_c(f_i, P_c)/\tau\right)$ from the denominator. Thus, Equation (II.33) is the naive adaptation of DCL to our prototypical segmentation setup. While using the original DCL [114] design is clearly beneficial with the image-wide representations z_i, for semantic segmentation it comes with the drawback that each pixel-embedding of the images is included in the contrastive term, which entails that when the positive association of the numerator lies within a large semantically coherent region, all pixel-embedding associations of this region will be present in the denominator. What this leads to is that semantically matching pixel-embeddings will be pushed apart from each other in the minimization of the loss. Put differently, the value of the numerator is increased and the value of the denominator decreased by means of increasing the mean cosine similarity $\sigma(\cdot)$ between f_i and $p_j \in P_c$ and decreasing this similarity for all other pixel-embeddings. Our pathway towards solving this problem lies in adjusting the denominator in a way to only include pixel-embeddings for which it can with high certainty be said that they do not match class c of the numerator. To this extent we enable the usage of any semantic cues in the form of different annotation types that may be available

for a given image in the batch. The altered normalization factor which decouples the denominator from the positive semantic class of the numerator can be stated the following way:

$$Z_{i,c} = \sum_{j=1}^{B\cdot H\cdot W} \sum_{k=1, k\doteq c \rightarrow k\notin \mathcal{A}_j}^{C} \exp\left(s_k(f_j, P_k)/\tau\right). \tag{II.34}$$

In this equation, the condition for including an association between a pixel-embedding and prototype set is adjusted. Firstly, if the precondition $k \doteq c$ of the implication is false, *i.e.* when $k \neq c$ all pixel-embedding associations to prototype sets other than the positive class c are included in the denominator (*ex falso quodlibet*). The second case is the critical part, the case when we are considering including a pixel-embedding which is associated to the positive class c, *i.e.* $k \doteq c$ is true. Then we have to check whether the semantic annotations for the j-th pixel indicate that it may belong to the class k (which is the positive class due to the fulfilled precondition). We formalize this via the notation of $\mathcal{A}_j$, which denotes the set of all possible classes for the pixel j based on the present annotations. If the pixel j lies in an unlabeled region, it could potentially be associated to all classes, therefore $\mathcal{A}_j$ is the set of all classes and $|\mathcal{A}_j| = C$. When the pixel j belongs to an image with an image-level label, A_j includes the set of all classes present in the image-level label. Further, if the image that the pixel j belongs to is associated with bounding boxes, $\mathcal{A}_j$ is the set of classes which have a bounding box overlapping with the coordinates of the pixel. Lastly, point- and mask annotations provide exact information about the class of the annotated pixel, thus, $\mathcal{A}_j$ for a pixel j with these annotations only includes the single annotated class, *i.e.* $|\mathcal{A}_j| = 1$. With this function $\mathcal{A}$ which returns the set of all possible classes at the pixel locations, the denominator in Equation (II.34) can be decoupled from the positive class completely. This happens in the right hand side of the implcation $k \doteq c \rightarrow k \notin \mathcal{A}_j$ in case the precodition is fulfilled. There, we only allow the inclusion of an association $s_k(f_j, P_k)$ iff the positive class is not in the set of possible classes A_j. With this strict exclusion of pixels-embeddings, we make

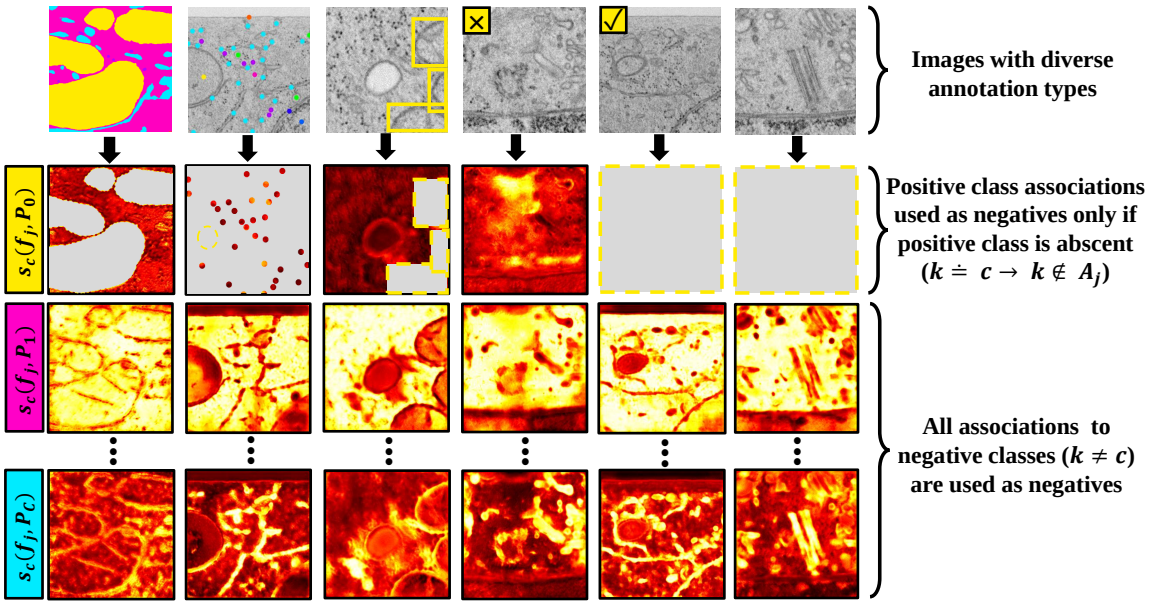


Figure 18: For our decoupled semantic prototypes, we disentangle the denominator of a contrastive term from the positive class based on diverse annotation types: masks, points, bounding boxes, image-level labels without and with positive class presence and unlabeled images. Yellow class (mitochondria) is the positive class, gray coloring in class score maps indicates regions of negatives omitted from the denominator.

sure that the denominator is completely decoupled from the positive semantic class. Here, we accept that some pixel-embeddings are excluded from the denominator even though their actual class might not have been the positive class, *e.g.* pixels within a bounding box might not belong to the box-class. On the other hand, in semantic segmentation it is possible to have bounding boxes completely filled with a rectangular class segment, thus this conservative over-exclusion of pixels is compatible with such edge-cases in segmentation and does not assume or even constrain the explicit class distributions within the images. In Figure 18 the process of decoupling the denominator of the contrastive term from a positive class by excluding associations $s_k(f_j, P_k)$ based on the possible present classes $\mathcal{A}_j$ is visualized.

By decoupling the contrastive term with respect to the positive class c, we ensure that only pixel-embedding-prototype pairs are pushed apart from the positive pair, that

with the certainty of a given annotation do not contain the positive class. Through our class-decoupling, all pixel-embeddings for a class c share the same negatives in the batch, which therefore only have to be computed once per class, and not per instance pairing, as in standard contrastive learning (*cf.* Section C.1 of the appendix).

6.2.2 Positive associations for decoupled semantic contrast

While Equation (II.34) provides a clear instruction on how to design the denominator and select the negatives for the contrastive term, the selection of positive associations in the numerator of Equation (II.29) has to be addressed next. For pixel-embeddings f_i which are associated to an exact class c, *e.g.* for mask- and point annotations the selection of the positive is natural: $s_c(f_i, P_c)$. Yet, reducing the set of positives to stem only from annotations that satisfy $|\mathcal{A}_i| = 1$ would severely reduce the set of positives and narrow the annotation types that could be used. In the following we outline how positives are chosen dependent on the annotation type present for a pixel-embedding f_i.

Pixel-wise masks For pixel-wise masks, we know exactly which class the pixel-embedding relates to. While directly taking $s_c(f_i, P_c)$ as the positive association is possible, we opt for integrating the notion of an instance-segment by pooling all $s_c(f_i, P_c)$ belonging to the same connected component in the mask. This pooled pixel-embedding-prototype pair is then considered as a positive of the contrastive term. We denote the positives obtained from mask annotations in a batch with respect to a class c as Ω_c^m.

Point annotations For pixel-embeddings f_i associated to a point annotation with class c, the embedding-prototype association $s_c(f_i, P_c)$ can be directly used in the numerator as positive pair. All point annotated positive pairs of class c within a batch make up the set Ω_c^p.

Image-level labels In case an image-level label contains the class c, a positive embedding f_i can be derived from the image via the idea of Multiple-Instance Learning (MIL) [159]. In MIL, the image can be considered as a bag of pixels, whereas the

image-level label is the label of the whole bag. To still derive class assignments for the individual pixels in the bag, a pooling function can be used to form a so called bag-level prediction. For our bag of pixel-embeddings, we therefore form a bag-level prediction by mean pooling all pixel-embedding associations to the positive class in the image: $\frac{1}{H \cdot W} \sum_{f_i \in F} s_c(f_i, P_c)$, where F is the whole embedding map for the given image. Ω_c^{im} is the set of positives in the batch obtained by pooling embedding maps of images with image-level labels containing the class c.

Bounding boxes In the work of Tian *et al.* [248], the authors leverage an implict property of bounding boxes which is that on each horizontal level and each vertical level of a bounding box, at least one pixel has to be associated to the box class c. With this property, we select the positive associations from bounding box annotations. Specifically, along all horizontal- and vertical lines of a box, we select the maximum embedding-prototype associations $s_c(f_i, P_c)$ and sum them up. By max-pooling along these lines in the box, we enforce that at least one pixel on each of these lines has to be associated to the box class c. This also respects semantic segmentation edge-cases where a segment is only one-pixel thin and diagonally oriented, which could lead to a large box with $w \cdot h$ pixels (box-width w and box-height h) but only $\sqrt{w^2 + h^2}$ pixels associated to the box class. The set of positives derived from bounding-boxes is Ω_c^b.

Unlabeled regions In case a complete image or regions within an image are not labeled, all these unlabeled pixels could potentially belong to any of the C classes. Thus, no positive candidates can be derived from them and they serve solely as providers of negatives in Equation (II.34).

Bringing together the full contrastive term with both the selected positives and decoupled negatives, the loss fuction can be written as:

$$L_{DSP} = \sum_{l \in \{m,b,p,im\}} \lambda_l \sum_{c=1}^{C} \sum_{f_i \in \Omega_c^l} L(f_i, c) \ , \tag{II.35}$$

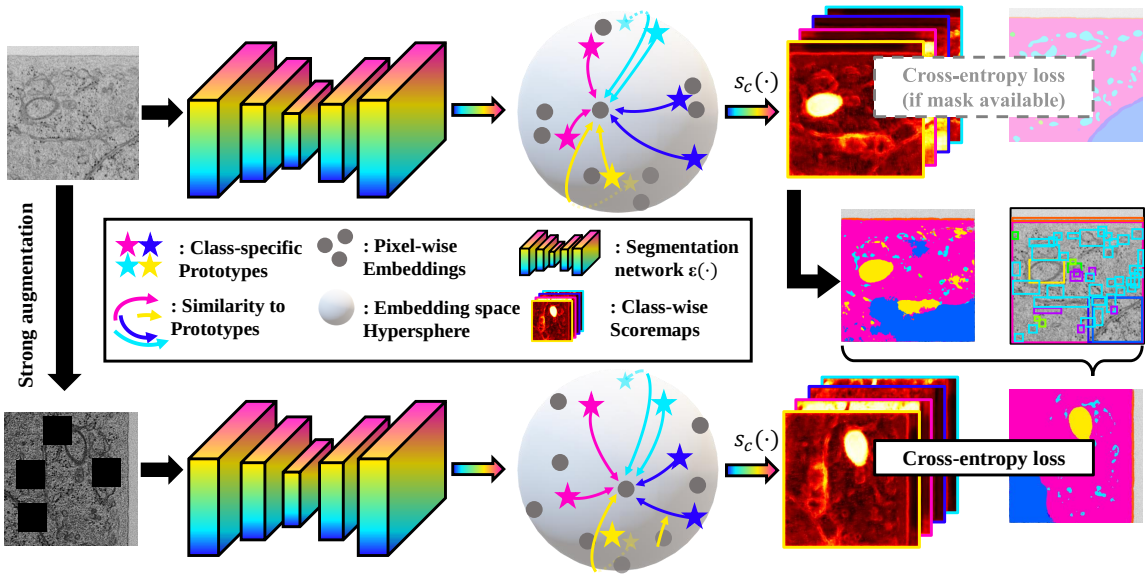


Figure 19: Our segmentation architecture which embeds each pixel of an input image into an embedding space and associates it with learned prototype vectors to obtain semantic predictions. On the right, we integrate the idea of pseudo-label filtering in order to refine self-inferred pseudo-labels with the available weak annotations.

where λ_l are weighting factors for the individual annotation type-based loss components. Details for the normalization of annotation type specific components can be found in Section C.1 of the appendix.

6.2.3 Pseudo-label filtering

In Section 4.2.2, we utilized a naive way to introduce weak annotations into the training of segmentation models, namely the notion of filtering pseudo-labels based on the weak annotations. We adapt the idea of forming a hard pixel-wise target for weakly annotated images using the network itself and then making this prediction coherent with the information given in either the image-level label, bounding boxes or point annotations. This is done, by altering the pre-softmax scores as follows:

1. **Image-level labels**: All predictions for class c are set to $-\infty$ if c is abscent from the image-level label.

2. **Bounding boxes**: At locations outside of all boxes with box-class c set the predictions for class c to $-\infty$.

3. **Point annotations**: At locations where point annotations are present, set the prediction to the point-class.

After filtering the predictions of the network in this fashion, a pseudo-label can be derived by computing the arg max as described in Section 4.2.2, (II.10). Specifically, we obtain pixel-wise predictions from a slightly augmented version of the input image and filter it for a more robust pseudo-label. As done frequently in semi-supervised literature [102, 34] we use this pseudo-label as target for a second prediction which is based on a strongly augmented version of the same image. This is similar in nature to the volume processing in Section 5.2.2. Based on the filtered pseudo-labels for weakly annotated images and standard pseudo-labels for unlabeled images, we subsequently compute a pixel-wise cross-entropy loss, which we refer to as L_{PLF} and display in Figure 19.

To enrich the baselines for our scenario of learning from mixed annotations, we add L_{PLF} to strong semi-supervised baselines making them semi-weakly trainable.

6.2.4 Decoupled prototypical nets for semi-weak supervision

With the altered architecture to segment images based on pixel-embeddings and learned semantic prototypes, the decoupled contrastive loss as well as the paradigm of strongly- and weakly augmented pseudo-label filtering, we can put together the complete training strategy. It consists of a standard pixel-wise cross-entropy loss L_{CE} (compare Equation (II.6)) for images where masks are available, our contrastive loss L_{DSP} for images with arbitrary annotation types which we apply on both weakly- and strongly augmented images, as well as the pseudo-label filtering loss L_{PLF} for augmentation invariance and additional integration of weak labels in training. This

leads to the complete loss function for semi-weakly supervised training of our *Decoupled Semantic Prototypes* for expert-centric segmentation:

$$L_{total} = L_{CE} + L_{PLF} + L_{DSP}. \tag{II.36}$$

6.3 Annotation compression ratio for semi-weak evaluation

The design of semi-supervised segmentation experiments is oftentimes done by arbitrary definitions of mask-annotated portions with respect to the whole training set, *e.g.* 5%, 10%, 20%. In previous chapters, we started investigating semi-weakly supervised segmentation by looking into the extreme case, *i.e.* using only one example per class and then subsequently doubling them in order to gain insight into the performance progression in these extreme scenarios. While both pathways give valuable insights into the behavior of segmentation algorithms with few annotations, *i.e.* their *annotation-efficiency*, a more systematic way of measuring this property is desirable. When we are given a dataset which is fully annotated with a *base annotation type* and we train an algorithm with a small annotation-portion of the whole annotation set, the algorithm can be thought of as compressing these annotations. With the perspective of compressing annotations when using fewer of them, we define the *Annotation Compression Ratio (ACR)* with regard to a base annotation type, which in our case of segmentation is pixel-wise annotation:

$$\text{ACR} = \frac{\#\text{ total base annotations}}{\#\text{ used base annotations}} \ . \tag{II.37}$$

This describes the degree of compression of the full annotations of a dataset, which can be used to see how different algorithms perform at different compression ratios. When training a segmentation algorithm with half of the masks in a dataset, it is trained at an ACR = 2, this notation is also common in the field of neural network pruning [249]. To make measuring the progression of algorithms with respect to the needed annotations at train time more systematic, we propose to successively

increase the ACR in an exponential fashion, *i.e.* $1, 2, 4, 8, 16, \ldots$ which respects that the accuracy as a function of the amount of annotations is generally regarded to follow power laws [250]. This way of probing the annotation efficiency at exponentially increasing ACRs equates to successively cutting the amount of annotations at training time in half. Instead of cutting mask annotations completely and training in a semi-supervised fashion, what we intend to do is to substitute them with weak annotations to end up at semi-weakly supervised training scenarios.

6.4 Experiments and results

In the next section, we present the experimental setup in which we test our semi-weakly *Decoupled Semantic Prototypes* method towards its *annotation-efficiency* via rigorous evaluation protocols leveraging the notion of the *Annotation Compression Ratio*. We present quantitative and qualitative results on an expert-driven test-bed for cell organelle segmentation within focused ion beam electron microscopy images. Afterwards, we discuss the insights from our expert-centric segmentation solution and how its experimental results relate to the research questions of this chapter.

6.4.1 Datasets

As dataset collection to benchmark semi-weakly supervised semantic segmentation approaches, we select the challenging OpenOrganelle data by Heinrich *et al.* [9]. In this collection, there are several individual large electron microscopy volumes, which each contain a whole imaged cell. Each such volume can be regarded as a dataset, where we put our focus on the datasets *HELA-2*, *HELA-3*, *JURKAT-1* and *MACROPHAGE-2* due to their difficulty and diversity. Example images showcasing the variability among different electron microscopy datasets is displayed in Figure 20. With the excessive size of these volumes obtained by focused ion beam scanning electron microscopes (FIB-SEM), they are only annotated in manageable sub-volumes. We use these annotated sub-volumes and extract small 2D slices from them to train our 2D segmentation models using the pixel-wise annotations which

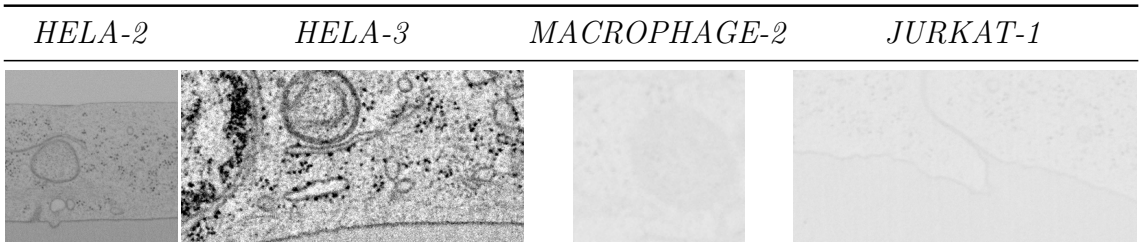


Figure 20: The cell organelle electron microscopy collection OpenOrganelle by Heinrich *et al.* [9] is highly variable regarding image properties across datasets.

cover very detailed cell organelle classes. In order to follow a strict protocol, we create cross-validation splits in a way such that training, validation and testing contain distinct sets of these sub-volumes. Further, we have to ensure, that in all train-, val-, test-sets, all classes are present. To have enough classes which are present in three different sub-volumes for this cross-validation setup, we merge classes in a biologically coherent fashion, *e.g. mitochondria*, *mitochondria membrane* and *mitochondria DNA* are merged into a single *mitochondria* class. After this merging process via a biologically motivated class hierarchy [9], we have 11 cell organelle classes for the *HELA-2* dataset, 10 for *HELA-3*, 8 for *JURKAT-1* as well as *MACROPAHGE-2* which are present in at least three sub-volumes. Classes which occur in less than three sub-volumes after merging are omitted as they do not fulfill the cross-validation requirement. In order to obtain multiple cross-validation splits, we shuffle the sub-volumes of a dataset and randomly distribute them into train-, val-, test-sets with the restriction, that all three sets contain all classes. As in previous chapters, we enumerate all slices within the training set for a reproducible selection of arbitrary portions of their annotations. This is done similarly to what we described in Section 4.3.2. For the largest dataset, namely *HELA-2*, we create a total of 10 cross-validation splits, for the remaining three datasets we use 5 to ease the computational requirements.

6.4.2 Evaluation protocol

In our experiments, we successively and exponentially reduce the amount of mask annotations used to train the different segmentation algorithms, *i.e.* we consecutively

double the ACR from 1 to 64, which equates to going from 100% to circa 1.6% of mask annotations. While reducing costly mask annotations, in our first stack of experimental setups, we substitute the omitted masks with either image-level labels or bounding boxes or point annotations. What we are equally as interested in is the scenario, when a model is trained with a diverse mix of annotations, covering: masks $\mathcal{M}$, boxes $\mathcal{B}$, points $\mathcal{P}$, image-level labels $\mathcal{I}$ and unlabeled images $\mathcal{U}$. To this end, we also successively increase the ACR as before, but instead of substituting masks with a single annotation type, we uniformly distribute all remaining annotation variants ($\mathcal{B}$, $\mathcal{P}$, $\mathcal{I}$, $\mathcal{U}$) among the images which are not associated to a pixel-wise annotation anymore. As an example for this mixed supervision scenario, with an $ACR = 2$, the models are trained with 50% pixel-wise masks, 12.5% unlabeled, 12.5% image-level labels, 12.5% point annotations, 12.5% bounding boxes.

To evaluate the efficacy, we follow the common procedure for segmentation models and infer the class-prediction P for all pixels in all testing images with a given model and calculate the DICE coefficient (or F1 score) using the ground-truth G:

$$\text{DICE}(P, G) = \frac{2|P \cap G|}{|P| + |G|} \tag{II.38}$$

To evaluate the performance on cell organelle segmentation, we calculate the DICE for all C classes of the respective electron microscopy datasets:

$$\text{mDICE}(P, G) = \frac{1}{C} \sum_{c}^{C} \frac{2|P_c \cap G_c|}{|P_c| + |G_c|} \tag{II.39}$$

The mean DICE is computed by averaging these individual class DICE scores, here P_c and G_c refer to binary predictions and ground-truth related to class c. As we perform cross-validation with multiple folds the final measures which we display are the average mean DICE and the standard deviation over S splits (*i.e.*, each result

reflects S trained segmentation models):

$$\text{average mDICE} = \frac{1}{S}\sum_{s}^{S}\text{mDICE}(P^s, G^s) \tag{II.40}$$

Here, P^s and G^s are the predictions and ground-truth of the current split s. In the following results we generally write mDICE as shorthand for average mean DICE.

6.4.3 Implementation details

In our experimental setup, we adopt a consistent implementation of all methods employing the widely used Unet architecture proposed by Ronneberger *et al.* [44]. While our segmentation approach and the baselines can be applied to other segmentation architectures, we deliberately select Unets for their inherent stability, ensuring that side-effects such as the absence of intricately tuned learning rate warm-ups are minimized. This is especially important due to the large number of trained models we will obtain in our experiments, *e.g.* with 7 ACRs and 10 cross-validation splits for the *HELA-2* dataset at least 70 models are trained per segmentation approach per semi-weakly supervision scenario. All models undergo training using the AdamW optimizer [251] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. A fixed learning rate of $6e^{-5}$ is employed, along with a weight decay of 0.01 and Xavier initialization [229] for the network weights. The training process is conducted on a multi-GPU setup consisting of four NVIDIA A100-40 GPUs, each with a memory capacity of 40 GB. Training is performed for 100 epochs on each split of the dataset. During training, validation is carried out every 10 epochs, and the model with the best validation performance is evaluated on the corresponding test set after completion of training. To accommodate the varying memory requirements of different training methods, the batch size is set to the maximum possible value within the method's memory constraints, ranging between 16 and 28. To ensure compatibility with batch processing, it is necessary for the inputs to have equal sizes. However, the sub-volumes of the investigated datasets

occasionally exhibit varying image width and height. We address this issue by zero-padding all images to match the respective maximum size. For weak augmentation, a set of transformations is applied, including horizontal and vertical flipping, rotations by $0°, 90°, 180°, 270°$, as well as brightness, contrast, saturation, and hue jittering, each with a factor of 0.2. To further augment the data with strong augmentations in addition to flipping and rotation, a jitter factor of 0.4 is applied, and CutOut [240] is employed up to nine times, removing small regions from the image.

6.4.4 Competing approaches

Basic Unet [44]: As a lower baseline, the shared architecture of all methods, the Unet architecture, is trained with only the pixel-wise annotated masks that are available in the respective experimental setup. The trained models serve as initialization for all remaining approaches.

Pseudo-label [99]: A common baseline for semi-supervised settings is training the network with pixel-wise annotations in a standard fashion and for unlabeled data, self-infer a pseudo-label and utilize it as target for the unlabeled example. In this baseline, we add the notion of pseudo-label filtering from Section 6.2.3, enabling it to use weak annotations as well.

Con2R [34]: As a semi-weakly designed method, we test the *Con2R* approach from chapter 5 on 2D cell organelle segmentation instead of 3D volume processing. Therefore, we adjust the receptive volume size $\mathcal{R}$ to a two dimensional 16×16 size. With some tuning, we noticed that the performance of this approach is positively affected when sampling the query- and neighbor sets for strongly- and weakly augmented versions of the input instead of just using the strongly augmented branch. Further, the semantic consistency constraint can be altered to also leverage pseudo-label-filtered predictions in case weak labels are available. For this we apply the filtering rules as well as a softmax normalization, which then serves as input to the semantic consistency computations.

FixMatch [102]: As a strong baseline, we adapt FixMatch from semi-supervised

classification to segmentation and further make it semi-weakly trainable by adding pseudo-label filtering as done for the prior two baselines.

Classification branch [186]: When we train models semi-weakly using pixel-wise annotations and image-level labels (scenario $\mathcal{M} + \mathcal{I}$), the approach of Mlynarski *et al.* fits as baseline, as they train an auxiliary classification branch of the segmentation network to integrate image-level labels conveniently into the training procedure. With this alteration, the Unet is altered towards a dual-head architecture with a segmentation and a classification output-head.

Euclidean/Geodesic point branch: We also investigate the scenario of training with pixel-wise annotations as well as point annotations (scenario $\mathcal{M} + \mathcal{P}$). For this more exotic supervision variant, we are not aware of existing techniques. Therefore, we take inspiration from interactive segmentation [252] where point cues are commonly used via distance maps. Coherently to the previous dual-head architecture, we design a semi-weakly supervised baseline, which regresses either the point annotation-based euclidean- or geodesic distance maps in a second output branch.

Box loss: For training with masks and bounding boxes (scenario $\mathcal{M} + \mathcal{B}$), an approach based on the box-based loss of Tian *et al.* [248] is used. Other mask- and box supervised approaches often rely on priors [157, 154] which tend to hold for natural images but not for expert-centric domains such as electron microscopy imaging.

DSP (Ours): To calculate pixel-embeddings, we modify the Unet architecture by replacing its final output convolution layer. We replace it by a sequence of operations, including batch normalization [222], 1×1 convolution with 64 kernels, LeakyReLU activation, and a final 1×1 convolution with 64 kernels. Through this replacement, at the end of the network, we obtain $D = 64$ dimensional embeddings. For each class, we utilize $|P_c| = 5$ learned prototypes and a temperature parameter $\tau = 0.05$, ensuring a proper scaling of the similarities between embeddings and prototypes. Additionally, we assign weights to the annotation type-specific components of the loss as $\lambda_m, \lambda_b, \lambda_p, \lambda_{im}$, with each weight set to 0.1.

6.4.5 Hyper-parameter sensitivity studies

Before carrying out the main experiments, we investigate the sensitivity of *DSP*'s hyper-parameters towards the performance as measured in mDICE. We carried out these experiments on the first split of the *HELA-2* dataset at an ACR of 8 and report the validation performance in Table 11. In the first batch of experiments, we investigate how the weighting factors λ affect the segmentation efficacy. We see that moving from not using L_{DSP} with $\lambda_m, \lambda_b, \lambda_p, \lambda_{im} = 0$ to an equal weighting of 1 improves the results considerably. Choosing a weighting which roughly leads to similar loss values for the individual components in row three as well as choosing a too low weight of 0.01 for the contrastive losses in row five led to worse performance. Best results of 59.5% are achieved by a simple equal weighting with a factor of $\lambda_m, \lambda_b, \lambda_p, \lambda_{im} = 0.1$. In the second group of ablation experiments, we alter the softmax temperature parameter τ, which we identify as an important hyperparameter, largely affecting the performance. Here, the original temperature of 0.05 from the previous batch of experiments remain the best results. The architectural design with our decoupled semantic prototypes allows for an arbitrary number of prototypes representing an individual class. In the third group we find, that between 1, 5 and 10 prototypes per class, 5 yield the best performance for the dataset at hand with 60.7% DICE. To showcase that the architectural alteration alone is not the singular factor leading to these improved results, we drop both L_{DSP}

L_{PLF}	λ_m	λ_b	λ_p	λ_{im}	τ	$\lvert P_c \rvert$	DICE
1.0	0.0	0.0	0.0	0.0	0.05	10	55.9%
1.0	1.0	1.0	1.0	1.0	0.05	10	58.4%
1.0	1.0	0.2	0.5	0.3	0.05	10	57.7%
1.0	0.1	0.1	0.1	0.1	0.05	10	59.5%
1.0	0.01	0.01	0.01	0.01	0.05	10	56.8%
1.0	0.1	0.1	0.1	0.1	1.0	10	41.9%
1.0	0.1	0.1	0.1	0.1	0.5	10	49.3%
1.0	0.1	0.1	0.1	0.1	0.01	10	58.6%
1.0	0.1	0.1	0.1	0.1	0.005	10	59.4%
1.0	0.1	0.1	0.1	0.1	0.05	1	59.0%
1.0	0.1	0.1	0.1	0.1	0.05	5	**60.7%**
0.0	0.0	0.0	0.0	0.0	0.05	5	48.9%

Table 11: Hyper-parameter sensitivity study for *DSP* on the first split of the *HELA-2* dataset. Training is done in the mixed supervision scenario at an ACR of 8. The configuration of $\lambda_m, \lambda_b, \lambda_p, \lambda_{im} = 0$ equates to not using L_{DSP}.

and the pseudo-label filtering loss L_{PLF} in the last row, where we can clearly observe a deterioration in performance to 48.9% DICE. An ablation study for the baseline methods can be found in Section C.5 of the appendix.

6.4.6 Quantitative results

Next, we evaluate all presented baseline methods on the task of semi-weakly supervised cell organelle segmentation. First, in Figure 21 we will show the results of all algorithms trained on the *HELA-2* dataset with (a) masks and image-level labels ($\mathcal{M}+\mathcal{I}$), (b) masks and bounding boxes ($\mathcal{M}+\mathcal{B}$), (c) masks with point annotations ($\mathcal{M}+\mathcal{P}$) and (d) mixed supervision with: masks, image-level labels, boxes, points and unlabeled data ($\mathcal{M}+\mathcal{I}+\mathcal{B}+\mathcal{P}+\mathcal{U}$). And finally, we show the generalization capability of our method on the datasets *HELA-3*, *MACROPHAGE-2* and *JURKAT-1* which we all investigate in the mixed supervision scenario in Figure 22. In order to keep the number of experiments manageable, for all mixed supervision scenarios, we only evaluate the lower baseline, our method *DSP* and the best baseline method from the experiments with pairs of annotation types (a) – (c). We present all results in graphs which plot the performance against the *Annotation Compression Ratio*, while the corresponding numerical values are available in Section C.2 of the appendix.

Supervision $\mathcal{M}+\mathcal{I}$: The first thing that can be investigated is the lower baseline, *i.e.* the Unet trained with merely the available masks at the given ACR. The basic Unet is always displayed in black and stays similar across experiments in Figure 21. As expected, reducing the amount of pixel-wise annotations also reduces the segmentation performance with the steepest decline happening between the ACRs 8 and 32, where it declines from 43.6% to 24.6% mDICE, an absolute drop of -19%. At an ACR of 64, *i.e.* with merely 36 pixel-wise annotations left, the basic Unet reaches 20.2% mDICE, which is better than a random baseline at 5.1% or predicting the most frequent class at 6.1% but is an overall very poor segmentation performance. In Figure 21a we can observe, that the integration of image-level information as done in Con2R, Pseudo-label and the Classification branch model certainly helps in slowing

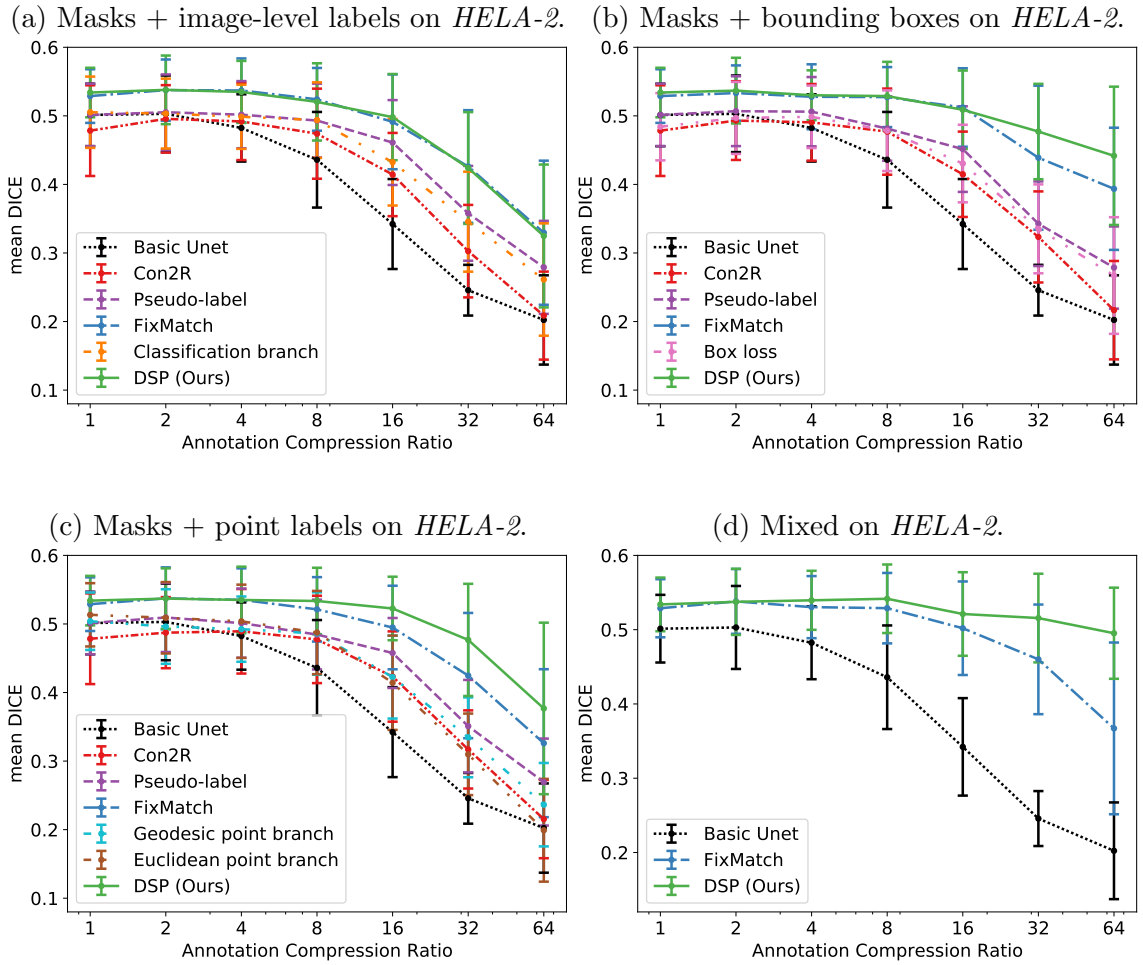


Figure 21: Graphs showing the performance in average mean DICE along 10 cross-validation splits with standard deviation. Experiments of semi-weakly supervised training on *HELA-2* for different supervision combinations (noted above the graphs). Performance reported as function of amount of masks present during training.

the degradation of the segmentation performance. Especially, the simple Pseudo-label method, which we augmented with the idea of pseudo-label filtering performs quite reasonably up to an ACR of 16, dropping only by -4% absolute mDICE as compared to the fully supervised 50.1%. In the extreme case scenario, the Pseudo-label method improves the basic Unet results by $+7.7\%$ absolute mDICE, showing

the value of adding image-level labels into the training, when only very few pixel-wise annotations are present. The two best performing models in Figure 21a are FixMatch with the pseudo-label filtering alteration and our *DSP* method. Directly from the start, *i.e.* at an ACR of 1, these two methods achieve a higher mDICE performance than all competing methods, reaching up to 53.4% for *DSP* and 52.9% for FixMatch. This notable positive offset in performance might stem from the addition of weak- and strong augmentations and suitable ways to integrate them for a certain degree of augmentation invariance. Comparing FixMatch and *DSP* in this scenario with masks and image-level labels, they lead to a quite similar performance.

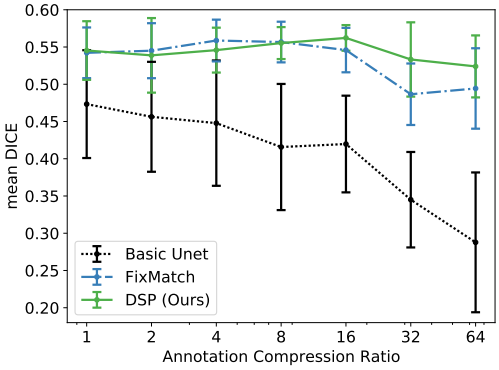
Supervision $\mathcal{M} + \mathcal{B}$: In Figure 21b, we can observe a similar behavior of the methods: the Unet baseline deteriorates fastest, in the middle ground Con2R, Pseudo-label and the Box-loss baselines perform slightly better, while FixMatch and *DSP* work best. Looking into the especially annotation scarce scenarios with ACRs of 32 and 64, the benefit of contrastive modeling in *DSP* is clearly evident, as it outperforms FixMatch by +3.8% and +4.8% mDICE respectively. This strong performance, which is specifically present when very few masks are available is a valuable property for expert-centric applications which typically operate in this territory.

Supervision $\mathcal{M} + \mathcal{P}$: When conducting training with mask and point supervision, in Figure 21c, the difference between the two front runners FixMatch and our *DSP* method becomes more pronounced. Our method outperforms FixMatch by +1.3%, +2.8%, +5.2%, +5.1% mDICE for ACRs from 8 to 64, again showing strong performance gaps especially in the scenarios where few masks are available. When comparing the performance of *DSP* between supervision configurations in Figure 21a, Figure 21b and Figure 21c, we can confirm, that aside mask annotations, bounding boxes provide the most information for training semantic segmentation models, with points following and image-level labels yielding the least hints for better training.

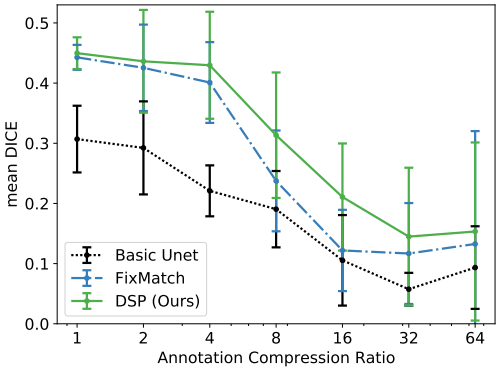
Supervision $\mathcal{M} + \mathcal{I} + \mathcal{B} + \mathcal{P} + \mathcal{U}$: Next, in Figure 21d, we investigate the setting, when models are faced with a diverse mix of annotation types, namely masks, boxes, points, image-level labels and entirely unlabeled images, which is our core motivation

in order to mature segmentation approaches towards flexible integration of heterogeneous semantic cues, making them more flexible towards expert-provided information. The first thing that can be noticed is, that the performance of FixMatch in the mixed supervision scenario is slightly better than when trained with masks and boxes with the exception of the extreme cases at ACRs of 16 and 64. If we compare the graph of FixMatch with *DSP*, we see that starting at an ACR of 4 the two graphs decouple from each other and *DSP* coming out on top, decreasing in performance much slower. The performance when training *DSP* with mixed annotation types is the best among all the different supervision scenarios we investigated in the very scarce annotation regimes of ACR 32 and 64. At an ACR of 64, with a performance of $49.5 \pm 6.1\%$ mDICE when trained with 36 mask annotations aside the mixed weaker annotation types, *DSP* comes close to the fully mask-supervised performance of the basic Unet which lies at $50.1 \pm 4.6\%$ using all 2321 mask annotations. When comparing the relative performance drop from ACR 1 to 64 (reduction of 98.4% in mask annotations), the basic Unet performance drops by -59.7%, FixMatch performance drops by -30.6% while *DSP* performs steadily and only

(a) Mixed on *HELA-3*.



(b) Mixed on *MACROPHAGE-2*.



(c) Mixed on *JURKAT-1*.

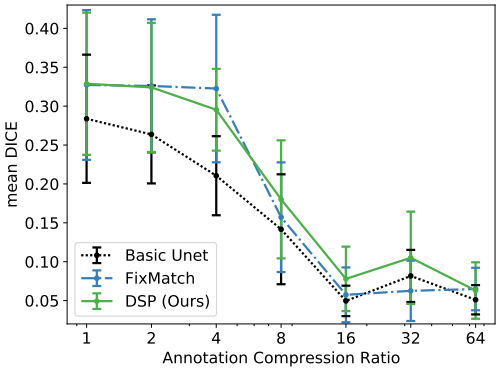


Figure 22: Results on *HELA-3*, *MACROPHAGE-2* and *JURKAT-1* over 5 splits in the mixed scenario.

drops by a relative -7.3%. In absolute terms, *DSP* achieves an absolute mDICE increase over the performance of FixMatch at ACR $= 64$ of $+12.8\%$. The results indicate that leveraging a diverse mix of annotation types as opposed to supplying only pairs of annotation types is not merely possible with *DSP*, but even has an advantageous effect on the segmentation efficacy. This might be due to *DSP* enabling a flexible integration of boxes which delimit entities and encode their spatial extent as well as point annotations which hint at important, central points on the semantic segments. In Figures 22a to 22c, we show that our *Decoupled Semantic Prototypes* method is also able to generalize to other datasets from the OpenOrganelle family when trained with the diverse mix of annotation types. For large ACRs, *i.e.* $16, 32, 64$, *DSP* is able to outperform other approaches, making best use of the few mask annotations and the remaining mix of annotation types. On the *JURKAT-1* dataset, all methods struggle in producing strong segmentation models when masks are scarce, which might go back to challenging image properties such as low contrast in the dataset.

A note on significance To make sure our *Decoupled Semantic Prototypes* bring about a significant improvement when masks are scarce on the *HELA-2* data as compared to prior art, we compare it using ASO [253] with a confidence level $\alpha = 0.05$, where we found the score distribution of our *DSP* based on ten cross-validation results to be stochastically dominant over FixMatch for the following scenarios:

$\mathcal{M} + \mathcal{B}$: ACR $= 32, 64$ ($\epsilon_{\text{min}} = 0.22, 0.11$)

$\mathcal{M} + \mathcal{P}$: ACR $= 16, 32, 64$ ($\epsilon_{\text{min}} = 0.12, 0.33, 0.02$)

$\mathcal{M} + \mathcal{B} + \mathcal{P} + \mathcal{I} + \mathcal{U}$: ACR $= 16, 32, 64$ ($\epsilon_{\text{min}} = 0.38, 0.06, 0.00$)

For reference, if $\epsilon_{\text{min}} < 0.5$, *DSP* is stochastically dominant over FixMatch in more cases than vice versa, as such it can be declared as superior. As noted earlier, in the $\mathcal{M} + \mathcal{I}$ scenario, both *DSP* and FixMatch perform comparably well, which is supported by our significance analysis. Yet, it has to be noted that we attribute FixMatch our pseudo-label filtering scheme to make it compatible with weak annotations. When comparing the basic semi-supervised FixMatch results with *DSP*

models trained with additional image-level labels, the score distribution of *DSP* is found to be stochastically dominant over FixMatch in the annotation scarce scenarios ACR $= 32, 64$ ($\epsilon_{\text{min}} = 0.48, 0.00$) which is to be expected due to the miss-match in annotations used. For this analysis we made use of the Almost Stochastic Order test [253, 254] as implemented by [255].

6.4.7 Qualitative results

After gathering insights from quantitative segmentation results across a broad set of baseline algorithms as well as our *Decoupled Semantic Prototypes*, we display the basic Unet results, the FixMatch and *DSP* segmentations trained in the mixed supervision scenario in Figure 23. Starting with the lower baseline, we see that the Unet trained exclusively with masks produces an acceptable organelle segmentation when supplied with sufficient pixel-wise annotations in the $ACR = 2, 4, 8$ scenarios. Starting with an ACR of 16, we can observe in the first row, that the **mitochondrium** class (yellow) is not captured anymore and the **cytosol** class (pink) starts leaking into the extra cellular space (transparent) at the top of the image. Further, the **cell nucleus class** (blue) is segmented very crudely and speckles of this class get scattered over the bottom half of the image. The Unet fails in accurately capturing the majority of cell organelle outlines for ACRs $16, 32, 64$, which is also confirmed in a second example in row four, where with 72 pixel-wise annotations (ACR $= 32$), it is merely capable to coarsely segment the majority classes **cytosol** and extra cellular space, while with 36 pixel-wise annotations (ACR $= 64$), **cytosol** is predicted for almost all pixels. For the extreme case of ACR $= 64$, FixMatch results look better as compared to the basic Unet, while it also mostly captures the majority **cytosol** class in row five, it is at least able to distinguish between the inside of the cell and the extra cellular regions. Yet, it has to be noted, that FixMatch was trained with an additional mix of weak annotations which through pseudo-label filtering and the training towards augmentation invariance is able to produce these slightly better results. Such differences to the basic Unet also show in row two,

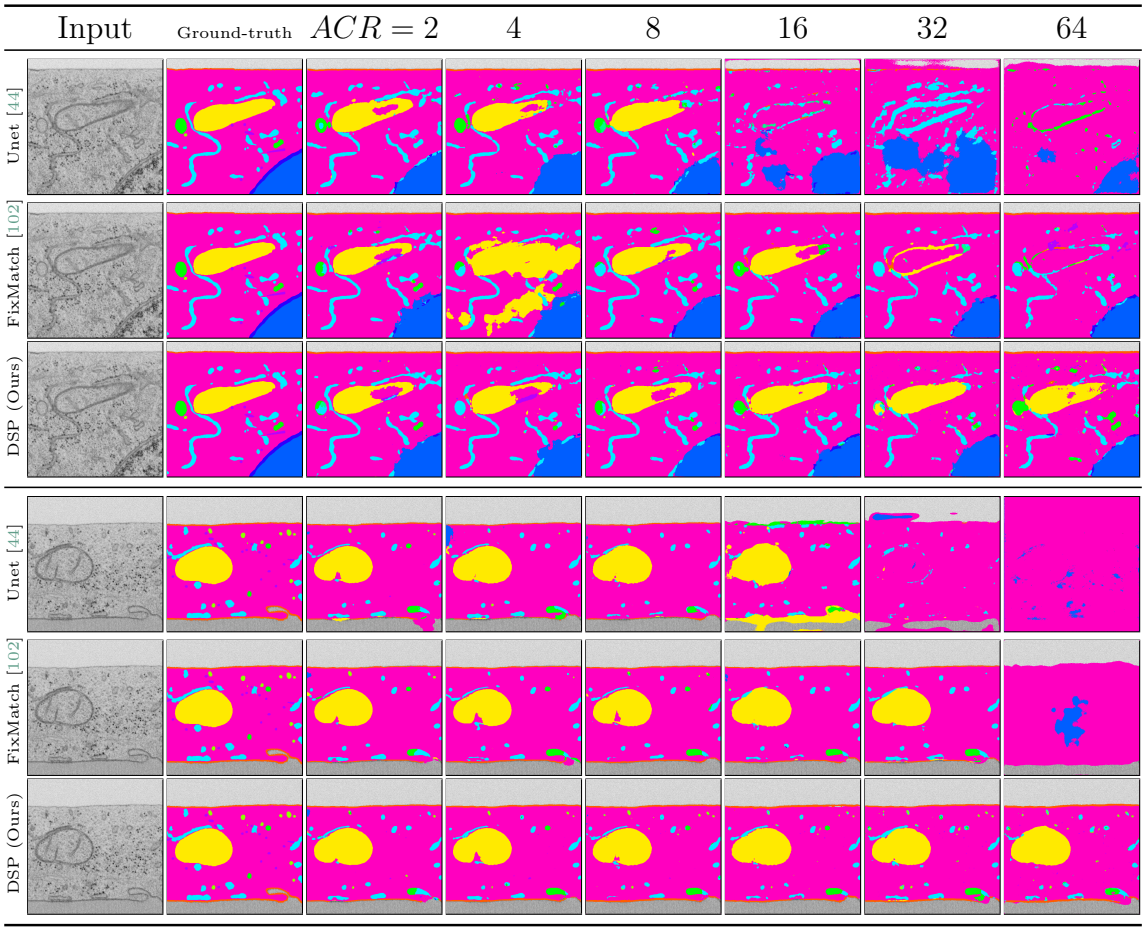


Figure 23: Qualitative segmentation results when training with a different amount of pixel-wise annotations plus a diverse set of annotation types: bounding boxes, points, image-level labels and unlabeled images. Two different images of cell organelles from the *HELA-2* dataset are segmented with a Unet, FixMatch and our *DSP*.

where for ACR = 64 nuanced contours of the cell organelles get captured better. When looking into the segmentation progression using more pixel-wise annotations, FixMatch produces perceptible organelle segmentations already for ACR = 16, while the basic Unet requires double the amount of annotations at ACR = 8 to produce visually matching results. For FixMatch one trained model, at ACR = 4 heavily over-segments the **mitochondrium** class for the given image which underlines the

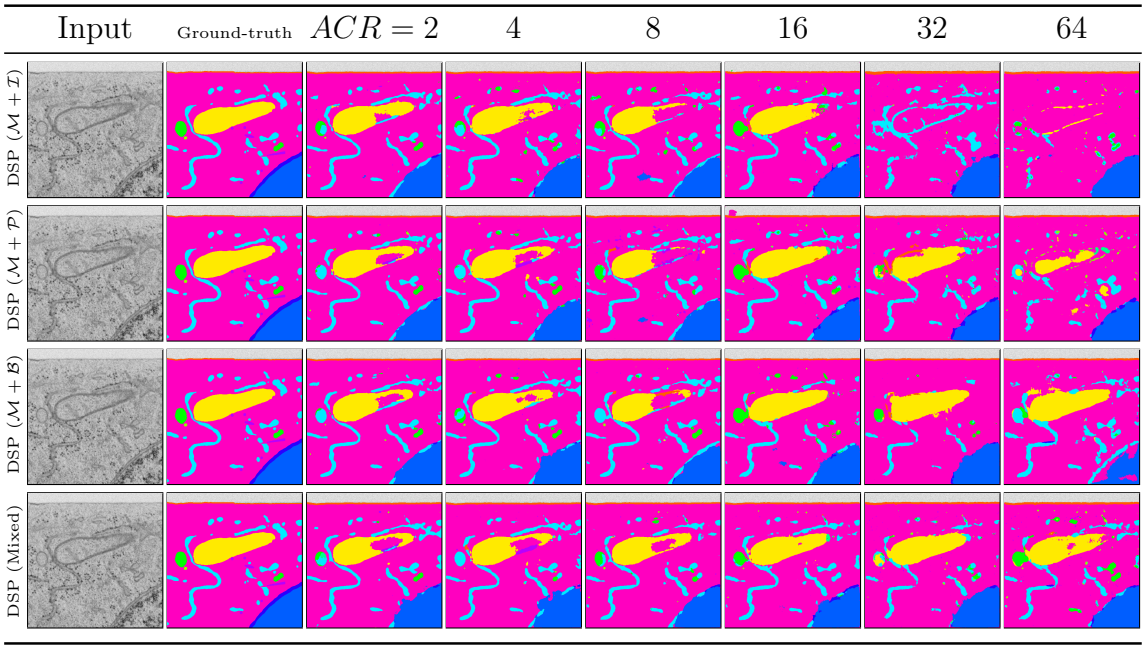


Figure 24: Difference of segmentation results when supervising *Decoupled Semantic Prototypes* with different combinations of annotation types on an image of the *HELA-2* cell organelle segmentation data.

need for multiple cross-validation splits in order to ease such effects happening in stochastic optimization, and rank the different segmentation approaches by their mean performance along multiple runs with differently split train-val-test sets. Moving towards our *DSP* approach, we can see, that it is able to much better utilize the weak annotation types to perform semantic segmentation in scenarios with very few pixel-wise annotations. For an ACR = 64, which equates to only 36 masks and the addition of boxes, points, image-level labels and unlabeled data through our contrastive decoupling visually improves noticeably upon the baseline results. Already there it is able to correctly capture the **mitochondrium** class, with only a few miss-classifications within the segment, but also the very thin **plasma membrane** (orange) gets captured accurately in this scenario both in row three and six. The improvement with more and more pixel-wise annotations can be observed in the segmentation of – for non experts – partially similar looking **endoplasmic reticulum**

(turquoise), **Multivesicular bodies** (green) and **vesicles** (grass green) classes. Yet, *DSP* struggles with the **microtubules** class (purple), which is hard to capture for all of the tested methods.

As we trained *DSP* in four different supervision scenarios, in Figure 24 we are able to display how training with masks and image-level labels ($\mathcal{M} + \mathcal{I}$), masks and points ($\mathcal{M} + \mathcal{P}$), masks and boxes ($\mathcal{M} + \mathcal{B}$) as well as mixed supervision ($\mathcal{M} + \mathcal{I} + \mathcal{P} + \mathcal{B} + \mathcal{U}$) affects its visual segmentation results. While training with masks and image-level labels helps in correcting miss-classified contours as compared to a basic Unet (compare Figure 23, row one), it does not help to identify coherent semantic regions as it does not supply added location information to the model. This changes when points are used, there, the model is able to identify larger regions lying on the actual organelles, yet with very few pixel-wise annotations, *e.g.* in the ACR = 64 scenario, if fails in correctly delineating the outlines of individual organelles. Using bounding boxes improves this problem visibly for the **mitochondrium** class but still produces miss-classifications within the **nucleus** (blue) and **nuclear envelope** (dark blue) for ACR = 64, which the model trained with a mix of all annotation types is able to correct.

6.5 Discussion

In the classical semantic segmentation pipeline, after the image collection phase, annotators supply a pixel-wise annotation to each of the images. When the availability of expert annotators is the critical resource, the limited time they have to annotate images has to be spent with caution. From an algorithmic standpoint, prior segmentation solutions were designed to work with pixel-wise masks exclusively, or added unlabeled data. Some work considered training segmentation models with two types of annotations such as masks and image-level labels. In our method *Decoupled Semantic Prototypes* we have shown that it is possible to go beyond these paradigms and integrate a wide diversity of annotation types into a unified training objective, namely masks, bounding boxes, points, image-level labels and unlabeled

images. This algorithmic solution has effects on the annotation phase of segmentation applications in expert-driven domains, as now, the expert annotators can spend their time more economically by flexibly providing different semantic cues which the subsequent segmentation algorithm can utilize to produce a more accurate trained model. Thus, it is now possible to opt for different degrees of granularity in annotations of a dataset which has either implications on the budget to be spent or the number of images which can be associated with semantic information. At the same time, such a flexible annotation phase has a large impact on how well the segmentation model performs. Instructing the experts to only annotate images with pixel-wise masks may lead to a sub-optimal allocation of the annotators time. When the expert's time is limited, it may be much more economical to gather the lions share of performance gain from only a few masks in conjunction with a diverse mix of annotation types for the majority of the images. This strategy yields datasets where the majority of images are linked to semantic information, while the classical strategy of only annotating pixel-wise covers only a small portion of images with semantic information, sacrificing diversity in the semantically annotated images. With our *DSP* algorithm, following this mixed-annotation strategy, with a reduction of 98.4% in mask annotations it achieves 98.8% of the performance as compared to a naive training algorithm which requires a dataset being completely annotated with dense, pixel-wise masks. When comparing our approach trained with masks, boxes, points, image-level labels and unlabeled data to trainings where only two of these annotation types are used, the diverse mix always outperforms the less diverse mix when very few masks are available. This, shows that in a scarce annotation scenario, as prevalent in expert-driven domains due to the expert's time constraints, opting for a mix of annotations can even be beneficial for the performance of a system. Lastly, being able to profit form diverse annotation types also enables an iterative annotation process, where annotations of different images are successively refined from unlabeled to image-level labels, to points and boxes and at the end to masks. Our flexible algorithm can make use of all images with their respective annotation

types at their different maturity stages. Our contributions summarize as follows:

Contribution 1:
We proposed an algorithm which is able to profit from annotation types as diverse as pixel-wise annotations, points, bounding boxes, image-level labels and unlabeled images to train a semantic segmentation model. With this, new ways of supervising segmentation models and thus, new ways for dataset annotation are made possible, which is a key component for applications and domains, where the annotation time is scarce due to the need for expert annotators.

Contribution 2:
With the notion of an exponentially increasing *Annotation Compression Ratio*, we offer a pathway to more systematically analyze semi-weakly supervised algorithms towards their efficient use of annotations with respect to their task accuracy. This way of designing an evaluation protocol is able to more precisely show how the accuracy of an algorithm progresses as it is supplied with fewer and fewer annotations of a specific type.

Contribution 3:
With our thorough evaluation and extensive experiments, we are able to quantify the effects of training different algorithms with different annotation type mixes, such as masks and image-level labels, masks and bounding boxes, masks and point annotations as well as models using masks, boxes, points, image-level labels and unlabeled data. This first analysis of its kind for semantic segmentation uncovers, that training with a broad diversity of annotation types is beneficial for the trained model's accuracy and superior to all other tested supervision scenarios, even though the annotation time spent is strictly smaller (under the assumption that boxes and points are more costly than image-level labels).

This concludes *Part II: Expert-centric Semi-weakly Supervised Semantic Segmentation* of this thesis. Next, concluding remarks regarding the impact of this work on the filed and exciting directions for continuing this research are discussed.

Part III

Concluding Remarks

7 Impact on the field

This thesis has advanced the research field of working with limited resources for the semantic segmentation of image data. Previous approaches either integrated unlabeled data besides pixel-wise annotations or exclusively worked with weak annotations and assumptions about the data distribution, while with our methodological contributions we enable networks to be supervised flexibly by diverse sets of annotation types or unlabeled images. We showed how to systematically probe into the annotation-efficiency of segmentation algorithms and gathered new information on the effects of supervising models with a broad mix of semantic cues. Here, we summarize the main contribution and opened pathways for exploring semi-weakly supervised segmentation solutions.

7.1 New research directions

Semi-weakly supervised volume segmentation: Learning from unlabeled volumes in conjunction with partially labeled volumes using sparse 2D labels was formalized in our work (Section 5.1.1) and introduced as a promising pathway towards easing the requirement of densely labeled volumetric data. The effect of establishing this research direction is two-fold: First, the developed solutions can be readily used to right now ease the requirement of dense volume annotations and profit from sparsely annotated as well as completely unlabeled volumes, making the annotation process easier for annotators. Secondly, with the clear establishment of the task of semi-weakly supervised volume segmentation, we set of a new promising way to view volume segmentation and to develop further solutions.

Semi-weakly supervised segmentation: Besides maturing the task of semi-weakly supervised segmentation to expert-driven domains such as retinal fluid segmentation in optical coherence tomography scans (Section 4.3) and cell organelle

segmentation in electron microscopy imaging (Section 6.4) we also formalized semi-weakly learning for semantic segmentation with diverse annotation type mixes, namely pixel-wise annotations, bounding boxes, point annotations, image-level labels and unlabeled data, where we directly propose a solution to work with such diverse sets of annotations. By establishing this semi-weakly supervised setting for training segmentation models, we set the stage for the exploration of segmentation solutions which are more flexible and accept diverse annotation types, which has direct implications on how images can be annotated, offering practitioners a broader set of possibilities to bring semantic segmentation solutions into new expert-centric domains.

7.2 New tools and insights in annotation scarce training

Establishment of clear protocols: Training segmentation models in scenarios where extremely few annotations are present comes with challenges in robustly measuring the results and establishing a performance ordering. For semi-weakly semantic segmentation on retinal fluid segmentation, brain tumor segmentation as well as cell organelle segmentation we establish evaluation protocols (Section 4.3.2, Section 5.3.2, Section 6.4.2) for a rigorous cross-validation reducing the effects of individual annotations and obtaining more reliable results which can be sidelined by suitable significance tests for deep neural networks. Our contribution here includes evaluation protocols utilizing the *Annotation Compression Ratio* (Section 6.3) to better quantify different degrees of precisely annotated supervision settings while varying the mix of annotation types used in training.

Insights from experiments: We included over $3,400$ network training results across all of our experiments which made it possible to gather considerable insights into how well semi-supervised and semi-weakly supervised segmentation solutions perform in scenarios where extremely few pixel-wise annotations are available (Section 4.3.6, Section 5.3.6, Section 6.4.6). This analysis further uncovered what effect different annotation mixes, or semi-weak supervision strategies have on the efficacy

of segmentation models and which strategies work better in these scenarios with a low amount of precise annotations, as prevalent in expert-centric applications.

7.3 Novel methods for semi-weakly supervised segmentation

At the center of this thesis, we propose novel methodological pathways to tackle semantic segmentation scenarios when costly, pixel-wise annotations are scarce. This perspective from the methodological side enables new possibilities to bring semantic segmentation solutions to applications, as it opens up how experts annotate image data, making the annotation process more flexible and enabling more cautious consideration of how to spend the expert-annotator's time to a maximum effect. Due to the wide variety of imagining techniques in expert-driven domains and thereby wildly different imagining properties, we have put special emphasis on designing algorithms which do not make harsh assumptions on the data distribution but are motivated from the standpoint of the task we want to solve, *i.e.* semantic segmentation, to enable better transferable methods as opposed to previous approaches. We proposed the *Mean-taught Deep Supervision* method (Section 4.2.3) including the *Multi-label Deep Supervision* loss function (Section 4.2.1), which enables networks to be trained better with noisy pseudo-labels through considerate, deep integration of these self-inferred semantic cues. Further, for training models with 2D partial labels in conjunction with completely unlabeled volumes, our *Contrastive Constrained Regularization* method (Section 5.2) can bring about performance improvements through considerations regarding common properties of the semantic segmentation task, *i.e.* smoothness properties both considering the semantic prediction space and a positional prior assumption (Section 5.2.1). Finally, we enable for the first time, the training with as diverse annotation types as masks, bounding boxes, points, image-level labels and unlabeled images for semantic segmentation with our *Decoupled Semantic Prototypes* method (Section 6.2). Our new, yet simple way of utilizing diverse annotation types enables a streamlined integration of supervision from differently granular labels, enabling expert-annotators to provide what they have the

time for as well as bringing in more junior experts in training for providing coarser semantic information. This advancement is sidelined by a more broadly formulated pseudo-label filtering notion for semantic segmentation (Section 6.2.3), which can easily be added to semi-supervised algorithms in order to make them semi-weakly trainable. Aside from bringing about a performance increase for semantic segmentation on the tested imaging data, our algorithms offer the possibility for flexible and annotation-efficient semantic segmentation in expert-driven domains.

8 Open questions for future work

With the proposed algorithms for bringing semantic segmentation better into scarce annotation environments, different pathways spring up which could be investigated in the future. These pathways include direct, natural extensions of the work presented here as well as the bigger picture when considering how a natural interaction between experts and an expert-centric learning system could look like in the future and which fundamental questions have to be answered first. As such, we continue the strain of thought from all previous chapters and outline promising research directions and goals to strive for in future endeavors to advance semantic segmentation in expert-driven domains.

8.1 A holistic view on annotation budgets

In essence, how many images can be supplied with annotations boils down to how large the annotation-time budget is, which may be constrained by monetary budgets, or by the availability of expert-annotators. With our *Annotation Compression Ratio*, we made a first step towards a more systematic evaluation of semi-supervised or semi-weakly supervised algorithms. Yet, pushing this idea further and considering not just a basic annotation type but integrating the costs of all annotation

types used via their associated average annotation-time expenditures into the *Annotation Compression Ratio* would open up a holistic view on the connection between annotation-time budgets and semantic segmentation efficacy. This could include investigations into measuring the time expenditure for one annotation of a specific type and how this varies between different datasets, imaging domains as well as the seniority of annotators in expert-driven domains.

8.2 A heuristic for dataset annotation

Equipped with an extended *Annotation Compression Ratio* and our proposed *Decoupled Semantic Prototypes* method which can handle diverse mixes of annotation types, all ingredients are there for a large scale investigation into the effect of training with different portions of annotation types. This is, at first glance, a natural extension of our work in chapter 6. Yet, when sampling the space of differently composed training sets in terms of number and types of annotations more densely and measuring it's effect on performance across a wide variety of datasets, a heuristic for annotating future datasets or guidelines for annotation become graspable. With a heuristic on how to go forward with the annotation of a set of unlabeled images, guidance can be given to practitioners, through which future segmentation endeavours could be greatly accelerated, while also spending the annotation budget in a much more economical, time-saving and results-driven fashion.

8.3 Heterogeneous training signals for flexible interaction

When humans teach other humans new skills or insights, it is an inherently multimodal process which may include any combination of verbal descriptions and interactions, pointing or gesturing, looking at few examples or reading in a text book supplied with figures and images. With our algorithms we were able to expand the realm of cues from which a single model can learn segmentation tasks to diverse subsets of masks, boxes, points, image-level labels and unlabeled data. By doing so, we

made the process of providing semantic information to the algorithm more flexible. Yet, there are a lot more ways to explain such information, including communication channels which are more natural to human nature such as the combination between speech, text and descriptive visuals. Extending training signals for a segmentation model to cover audio- or textual descriptions of a segmentation task at hand, learning from the gaze movement of a medical doctor when assessing a medical scan or extracting domain knowledge from highly technical textbooks may be a pathway to make the process of annotation or more generally of conveying a segmentation task even more flexible and natural. Probing into this field would require the investigation how segmentation tasks can be described via speech or text and unify a wide array of intricate training signals.

Part IV

Appendix

A Additional details for chapter 4

In this section, we show in more detail how we designed and tuned the baseline methods *IIC* and *MIL* as well as their deeply supervised versions from chapter 4.

A.1 IIC baseline

We integrate an additional information invariant over-clustering output-head on top of the Unet architecture besides the already present segmentation output-head and integrate the loss function as proposed in Ji *et al.* [230] leading to the IIC baseline. To investigate the performance of this model, we evaluate it on the validation scenario with full access to pixel-wise mask annotations and leverage the IIC loss for the IIC output-head and the standard cross-entropy loss for the segmentation output-head. The results are presented in Table A1 starting with the first line, which shows a standard Unet trained with all pixel-wise annotations on our evaluation setup for the RETOUCH retinal fluid segmentation dataset [8]. As IIC is trained with two forward passes per iteration, the second line provides results for a Unet with double the iterations to ensure that merely double the amount of processed images is not the deciding factor in IIC's performance. The next three lines in Table A1 show IIC's segmentation performance with an increasing number of clusters in the over-clustering output-head.

epochs	clusters	f	validation (mIoU)
100	–	–	62.42 ± 4.11
200	–	–	62.54 ± 3.88
100	5	f_4	63.42 ± 4.32
100	10	f_4	64.33 ± 2.84
100	20	f_4	64.63 ± 3.36
100*	10	$f_{\{0-4\}}$	$\mathbf{65.23 \pm 3.58}$

Table A1: Ablation study for models with IIC training as indicated by the column *clusters*, trained with full access to pixel-wise annotations. Smaller batch size (8 instead of 16) due to memory constraints indicated by the * symbol.

It is evident that adding more clusters increases the performance. Yet, more clusters (output-channels) also increase the memory footprint which is why we used 10 clusters for the experiments in chapter 4, in order to have a manageable memory consumption and faster training iterations. The overall best results for IIC are achieved when adding 10 over-clustering output-heads onto the feature maps $f_0, \ldots, f_4$ in the decoder. Yet, due to the added over-clustering output-heads, this Deeply Supervised IIC configuration leads to an even higher memory consumption which necessitated a decrease in batch size from 16 to 8 images. The neighborhood displacement hyperparameter (see [230] *Section 3.3 Implementation* for details) is always set to 5.
As the over-clustering output-heads are trained completely unsupervised, it is not clear beforehand on what regions in an image they activate. In Figure A1, we display the input OCT b-scan, the associated ground truth for retinal fluid segmentation as well as the activation in each channel of the over-clustering output-head by scaling the channel's values to a range of $[0, 1]$. Some of the channels show high activation in the different anatomical layers of the retina and also on the outside of the retina. Some approaches for retinal fluid segmentation integrate retinal layer segmentation explicitly into their training [256], thus, it is apparent that IIC's ability to learn this retinal anatomy leads to better segmentation as was seen in Table A1.

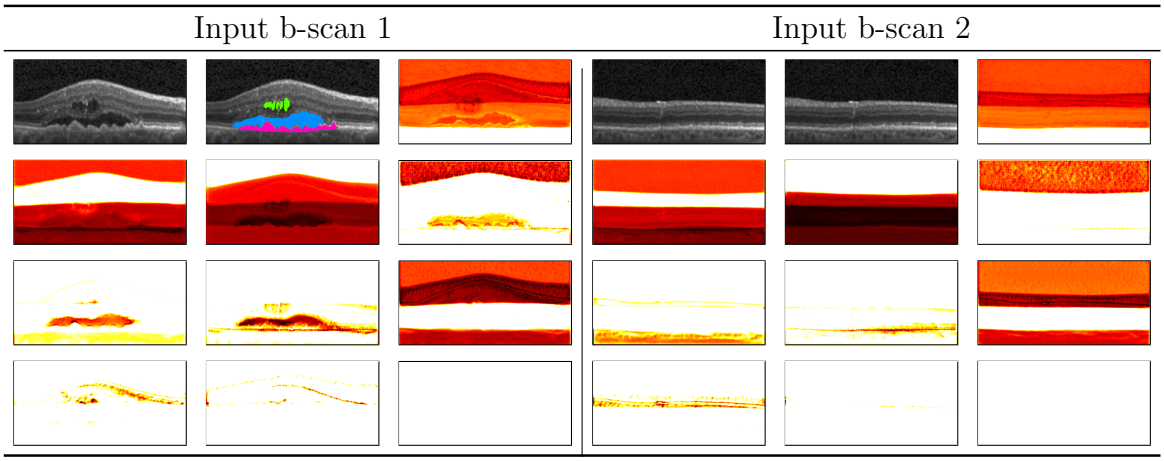


Figure A1: B-scans and their associated ground-truth annotation displayed alongside the feature scaled activation of 10 individual channels in the over-clustering output-heads of a trained IIC model. Left: b-scan of retina with fluid, right: healthy b-scan.

A.2 MIL baseline

For the MIL baseline model which spatially pools feature maps and applies a binary cross-entropy loss to enforce a multi-class image-level label as target, we investigate which feature map combination in the Unet decoder best fits to apply the MIL loss to. In Table A2, we see the results when training the MIL baseline models with 24 pixel-wise and additional image-level labels for the remaining images as outlined in our training protocol. Compared to the Unet which is trained only on pixel-wise masks in line one, all MIL-trained models successfully integrate image-level information, regardless of the feature map combination in the decoder the MIL loss is applied to.

Yet, applying it to all decoder feature maps $f_0, \ldots, f_4$ yields the best performance. This configuration serves as our *Deeply Supervised MIL* baseline. Next, we investigated which pooling function for the Multiple-Instance Learning loss leads to the best results in Table A3, considering the same 24 annotation plus image-level label scenario as in Table A2. First, we test max pooling, which was previously done in a completely weakly supervised segmentation setting by Pathak *et al.* [138]. Compared to the Unet trained on the masks only, using max pooling for the MIL loss in our Deeply Supervised MIL training only very marginally increased the performance, while the standard deviation almost doubles. When exchanging max pooling with average pooling,

$\mathcal{G}$	f	validation (mIoU)
–	–	46.84 ± 6.49
✓	f_4	49.48 ± 4.88
✓	$f_{\{3,4\}}$	50.13 ± 6.25
✓	$f_{\{2,3,4\}}$	51.81 ± 5.58
✓	$f_{\{1,2,3,4\}}$	51.47 ± 4.03
✓	$f_{\{0,1,2,3,4\}}$	$\mathbf{53.52 \pm 4.69}$
✓	$f_{\{0,1,2,3\}}$	52.85 ± 4.50
✓	$f_{\{0,1,2\}}$	51.81 ± 6.25
✓	$f_{\{0,1\}}$	51.68 ± 5.48
✓	f_0	50.23 ± 7.38

Table A2: Ablation study for models with MIL training and its deep integration into different feature maps f. First row indicates performance without image-level labels, second row *Baseline MIL*, using $f_{\{0,1,2,3,4\}}$ equates to *Deeply Supervised MIL*.

we get much better and more stable results as indicated by the last row. Therefore, we conclude that the average pooling MIL variant is more robust in optimization and leads to better results. This is in coherence with weakly supervised literature, where classifiers are often trained with global average pooling layers to extract coarse location cues in so called class activation maps [162]. The MIL baselines in chapter 4 are all trained with average pooling to aggregate features spatially.

$\mathcal{G}$	pooling function	validation (mIoU)
	–	46.84 ± 6.49
✓	max pooling	46.96 ± 12.73
✓	average pooling	$\mathbf{53.52 \pm 4.69}$

Table A3: Ablation results for pooling functions as used in the *Deeply Supervised MIL*-trained model.

B Additional details for chapter 5

In this section, we show in more detail how we designed and tuned some baseline methods, specifically, Pseudo-label, Uncertainty-aware Mean-Teacher as well as FixMatch and we show how the augmentation strategies used in chapter 5 were determined. It is described how baselines were adapted from 2D to 3D, which strong augmentations are suitable for training and which method specific hyper-parameters yield the best results to end up at strong baselines.

B.1 Pseudo-label

In implementing the Pseudo-label [99] baseline, we opted for an online pseudo-labeling procedure, where the unlabeled volumes are associated to pseudo-labels by on-the-fly predicting them within the training iteration using the current 3D Unet itself. This naive implementation led to diverging results, thus, we adapted it to an offline approach where the pseudo-labels are computed before training using the standard 3D Unet baseline which was trained on the annotated portions of the partial labels only. By integrating the loss normalization from [62] the segmentation results

got better. This normalization uses the loss on the partially labeled L_h regions and the loss on pseudo-labeled data L_p and weights them by:

$$\hat{L} = \frac{1}{1+\alpha}(L_h + \alpha \frac{\bar{L_h}}{\bar{L_p}} \cdot L_p) \ , \tag{IV.1}$$

where $\bar{L_h}$ and $\bar{L_p}$ denote the exponential moving averages over the two losses. The momentum parameter as specified in [62] is set to $\alpha = 4.0$, subsequently we minimize the loss $\hat{L}$.

B.2 Uncertainty-aware Mean-Teacher

We integrate the Uncertainty-aware Mean-Teacher [125] into the volumetric segmentation setting, by basing it on the 3D Unet architecture and adding the dropout layers needed for Monte-Carlo [242] sampling-based uncertainty calculations directly after the encoder as well as before the pixel-wise classification layer. We set the dropout probability coherently to the one in the original paper at 50%. Since we initialize all semi-supervised models with weights pre-trained on the available partial annotations (specifically, the standard 3D Unet models), we do not utilize the Gaussian scheduling or the successive up-weighting of the semi-supervised loss term as outlined in the training details of [125]. The Uncertainty-aware Mean-Teacher selects confident portions which determine the regions that are used for the consistency loss via the formula $U < U_{\max} \cdot \gamma$. In Table A4, we ablate which threshold value γ for the uncertainty values U works best on the validation set, which is $\gamma = 0.5$.

threshold	val mIoU
$\gamma = 0.1$	42.98 ± 5.06
$\gamma = 0.3$	42.96 ± 5.44
$\gamma = 0.5$	44.45 ± 4.06
$\gamma = 0.7$	41.98 ± 7.13

Table A4: Uncertainty-aware Mean-Teacher results with different thresholds.

A hyper-parameter for Monte-Carlo dropout is how often a single volume is forwarded through the 3D Unet to obtain multiple stochastic predictions for computing the voxel-wise uncertainty values U. In Table A5 we see that increasing the forwardpasses leads to a better validation performance, but also adds computational load. Therefore, we select 8 forward passes for computing U. In adapting Monte-Carlo dropout to volume segmentation, we investigate whether dropping out complete channels, *i.e.* 3D features or classical dropout works better in Table A6.

threshold	val mIoU
4 forward passes	44.16 ± 5.89
8 forward passes	44.45 ± 4.06

Table A5: Uncertainty-aware Mean-Teacher results with different numbers of stochastic forward passes in Monte-Carlo dropout.

threshold	val mIoU
classical dropout	44.45 ± 4.06
3D feature dropout	39.50 ± 5.10

Table A6: Uncertainty-aware Mean-Teacher results with different dropout variants.

B.3 FixMatch

As a strong baseline, we train FixMatch [102] models utilizing a standard cross-entropy loss to compare the predictions obtained from strongly augmented volumes with the pseudo-labels derived from weakly augmented volumes. Regarding weak flip augmentation strategies we conduct experiments in Table A7. From the analysis, employing a conservative weak augmentation scheme of only flipping volumes in the longitudinal- and vertical directions with a probability of 50%

horizontal	vertical	longitudinal	val mIoU
✓	✓	✓	46.05 ± 5.03
-	✓	✓	47.17 ± 4.13
✓	-	✓	46.35 ± 6.72
-	-	✓	45.94 ± 6.00
-	-	-	43.06 ± 6.32

Table A7: FixMatch validation results when changing the weak flip augmentation strategy.

leads to the best performance. As for the strong augmentation strategy, we evaluate photometric augmentations in Table A8 by adjusting brightness, gamma value and sharpness. The experiments indicate, that FixMatch with brightness- and sharpness perturbations using a magnitude sampled uniformly from $[0, 2]$ in addition to flipping as choice for strong augmentations work best on the validation split. Another possible addition to the strong augmentation branch is the CutOut [240] augmentation used in semi-supervised classification. Instead of cutting out image portions from 2D images, we cut out small volumes from the input to the strongly augmented prediction branch and ignore corresponding areas in the pseudo-labels. For 2D segmentation with FixMatch, CutOut was previously studied in [257], we evaluate the volumetric CutOut variant in Tab. A9. Cutting out large chunks of size $16 \times 16 \times 16$ produced the best validation results. Regarding the learning rate scheduling in training, the original publication included a cosine annealing strategy, yet, in our setting we

brightness	gamma	sharpness	val mIoU
✓	-	-	45.78 ± 7.75
-	✓	-	44.12 ± 8.14
-	-	✓	45.20 ± 8.06
✓	✓	-	28.38 ± 20.04
✓	-	✓	46.05 ± 5.03
-	✓	✓	42.47 ± 7.03
✓	✓	✓	23.12 ± 19.95

Table A8: FixMatch validation results when varying the photometric augmentation strategy.

CutOut size	val mIoU
none	47.17 ± 4.13
$4 \times 4 \times 4$	46.26 ± 5.57
$8 \times 8 \times 8$	45.72 ± 6.14
$16 \times 16 \times 16$	48.22 ± 5.65

Table A9: FixMatch results when using CutOut [240], we vary the cut out cube size.

learning rate	val mIoU
constant $lr = 0.01$	48.22 ± 5.65
cosine	45.27 ± 5.73

Table A10: FixMatch results with different learning rate schedules.

find that a simple constant learning rate produces better results in Tab. A10. Similarly, we find, in Tab. A11, that choosing a pseudo-label threshold of $\tau = 0.5$ for our volume segmentation task worked better than $\tau = 0.95$ which was used for the original classification task. Finally, we investigate the impact on the performance when down-weighting the loss for the unlabeled data by λ_u in Table A12, which is a common semi-supervised strategy [231] to manage a too pronounced impact of unlabeled data. We find that assigning an equal weight to the loss on labeled and pseudo-labeled data yields the best validation performance, which might be due to the fact, that we already balance the loss impact of partially labeled and unlabeled volumes by constructing balanced batches.

confidence threshold τ	val mIoU
$\tau = 0$	45.54 ± 6.72
$\tau = 0.2$	43.84 ± 6.50
$\tau = 0.5$	46.05 ± 5.03
$\tau = 0.7$	43.74 ± 8.23
$\tau = 0.95$	43.77 ± 7.78

Table A11: FixMatch results when tuning the pseudo-label confidence threshold τ.

weighting factor λ_u	val mIoU
$\lambda_u = 1.0$	46.05 ± 5.03
$\lambda_u = 0.5$	45.78 ± 6.13

Table A12: FixMatch results with different weighting of unlabeled examples.

B.4 Con2R voxel-embedding visualization

To get a qualitative grasp of what Con2R learns in the voxel-embedding branch, we visualize voxel-embeddings for the voxels of an individual slice taken from an OCT volume which we see in Figure A2. To visualize it, we forward the OCT volume through the trained network and obtain the voxel-embeddings from the output-head $\tau(\cdot)$.

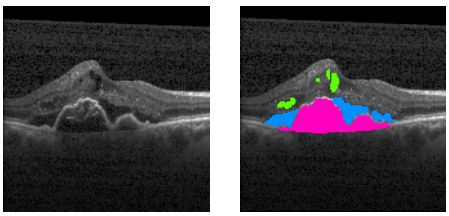


Figure A2: Left: Slice of input OCT volume, right: associated ground-truth.

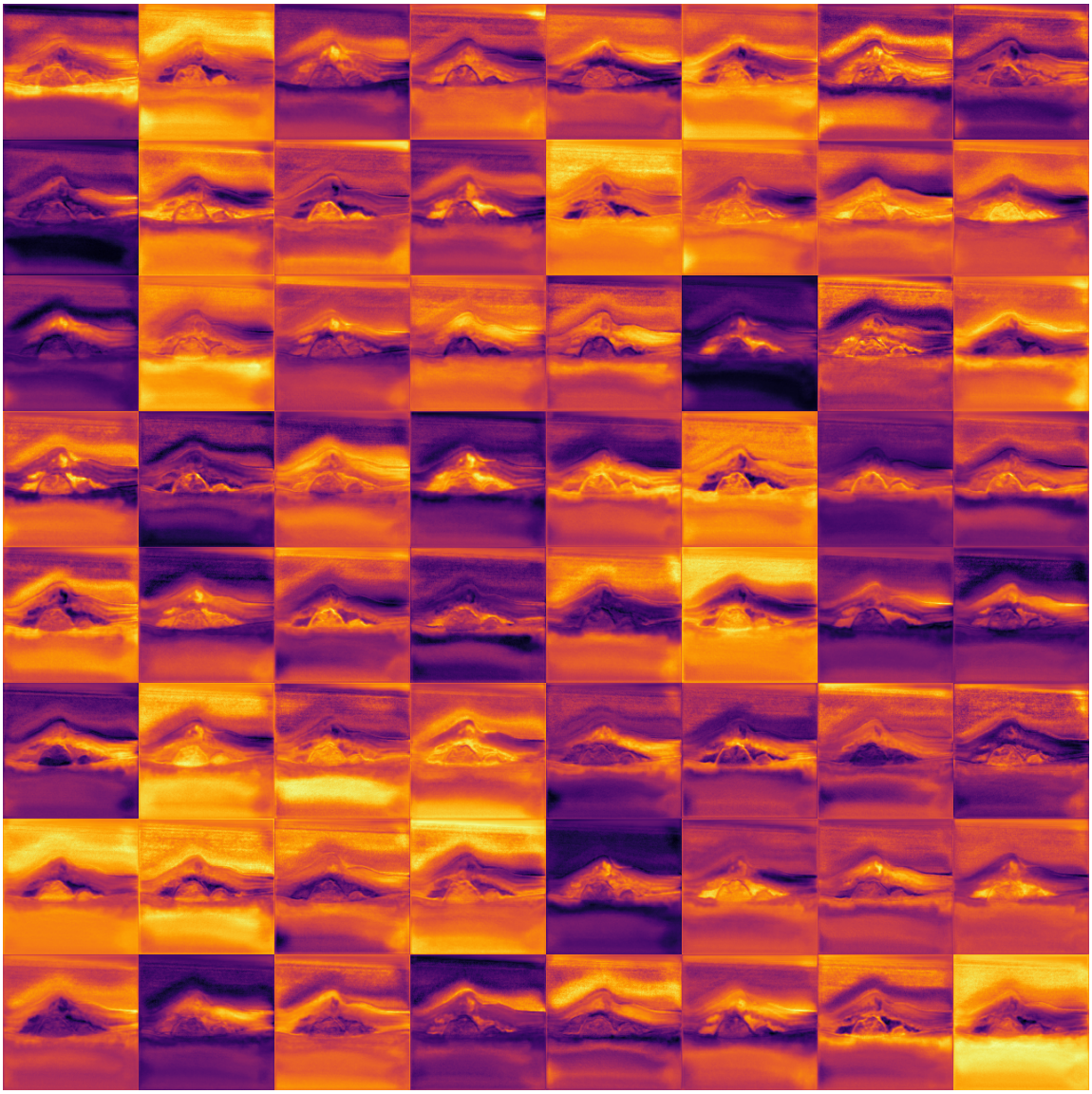


Figure A3: Individual voxel-embedding channels obtained from the output head $\tau(\cdot)$ and normalized by channel-wise feature-scaling for visualization purposes.

Then, we take each of the 64 dimensions of a single depth dimension (*i.e.* the corresponding depth of the slice in Figure A2) and visualize them all as 2D images in Figure A3 after normalizing their values. The embedding channels display very diverse and semantically relevant structures of the OCT slice, including retinal boundaries, fluid-types, the outside of the retina as well as the different layers of the retina.

C Additional details for chapter 6

C.1 Simplification and efficient computation of DSP's loss

For a more efficient implementation, we present, how the contrastive L_{DSP} loss term can be adjusted due to the fact that it decouples individual classes in the loss computation. In case of semantic segmentation this is especially important, as for each pixel individual contrastive negatives would have to be collected. Yet, via our decoupled design, negatives have to be computed only for each class instead of each pixel. Next, we outline how we implemented the final loss term of Equation (II.35) in a more efficient manner, with which we begin here:

$$L_{DSP} = \sum_{l \in \{m,b,p,im\}} \lambda_l \sum_{c=1}^{C} \sum_{f_i \in \Omega_c^l} L(f_i, c) \quad . \tag{IV.2}$$

In the last sum of this equation, for a class c and an annotation type l, the sum can be written, with the definition of $L(f_i, c)$ (Equation (II.32)), as:

$$\sum_{f_i \in \Omega_c^l} L(f_i, c) = \sum_{f_i \in \Omega_c^l} -\log \frac{\exp\left(s_c(f_i, P_c)/\tau\right)}{Z_{i,c}} \quad . \tag{IV.3}$$

Now, $Z_{i,c}$ from Equation (II.34) can be inserted into this formula, yielding our decoupled contrastive term:

$$Z_{i,c} = \sum_{j=1}^{B \cdot H \cdot W} \sum_{k=1, k \doteq c \rightarrow k \notin \mathcal{A}_j}^{C} \exp\left(s_k(f_j, P_k)/\tau\right) \quad . \tag{IV.4}$$

The requirement $k \doteq c \rightarrow k \notin \mathcal{A}_j$ in the second sum is designed to accept negatives either when $k \neq c$, *i.e.* it can be safely assumed that the pair is not related to the positive class c or when $k \doteq c$ and $c \notin \mathcal{A}_j$ is the case, meaning that the annotation for the given pair provides definitive information that it is not related to the positive class. Due to this decoupling of the denominator $Z_{i,c}$ becomes independent of f_i and only dependent on the current positive class c, so we write Z_c for short. By applying the logarithmic division law to Equation (IV.3), it can be simplified into:

$$\sum_{f_i \in \Omega_c^l} L(f_i, c) = \sum_{f_i \in \Omega_c^l} -(\log \exp\left(s_c(f_i, P_c)/\tau\right) - \log Z_c) \tag{IV.5}$$

$$= \sum_{f_i \in \Omega_c^l} -\left(s_c(f_i, P_c)/\tau - \log Z_c\right) \quad . \tag{IV.6}$$

Now that Z_c is independent from f_i it can be brought in front of the summation by multiplying it by the number of positives in the set Ω_c^l:

$$\sum_{f_i \in \Omega_c^l} L(f_i, c) = |\Omega_c^l| \cdot \log Z_c - \sum_{f_i \in \Omega_c^l} s_c(f_i, P_c)/\tau \quad . \tag{IV.7}$$

Scaling the negatives Z_c by a value > 1 produced large values for the loss, which was experimentally hard to handle. To counteract this, we re-scaled the loss by instead dividing it by $|\Omega_c^l|$, leading to:

$$L(c) = \log Z_c - \frac{1}{|\Omega_c^l|} \cdot \sum_{f_i \in \Omega_c^l} s_c(f_i, P_c)/\tau \quad . \tag{IV.8}$$

Integrating everything into our final loss, which is how we implement it, we obtain the loss function $L^{\star}_{DSP}$:

$$L^{\star}_{DSP} = \sum_{l \in \{m,b,p,im\}} \lambda_l \sum_{c=1}^{C} L(c) \cdot \delta(|\Omega^l_c| \neq 0) \ , \tag{IV.9}$$

here, the delta function $\delta(\cdot)$ is used to prevent division by zero which would happen in case $|\Omega^l_c| = 0$, *i.e.* when class c does not occur in the batch. We use this loss on both augmentation branches, the weakly- and the strongly augmented images.

C.2 Quantitative segmentation results in numerical form

Method	$\mathcal{I}$	$\mathcal{B}$	$\mathcal{P}$	$ACR = 1$	$ACR = 2$	$ACR = 4$	$ACR = 8$	$ACR = 16$	$ACR = 32$	$ACR = 64$
UNet	–	–	–	50.1 ± 4.6	50.3 ± 5.6	48.2 ± 4.9	43.6 ± 7.0	34.2 ± 6.6	24.6 ± 3.7	20.2 ± 6.5
CLS Branch [186]	✓	–	–	50.4 ± 5.5	50.3 ± 5.1	50.4 ± 4.6	47.3 ± 5.8	43.8 ± 6.4	34.6 ± 7.3	26.1 ± 8.2
Box Proj. [248]	–	✓	–	48.4 ± 4.9	49.7 ± 5.2	49.9 ± 4.6	47.8 ± 5.9	43.0 ± 5.7	33.5 ± 6.5	26.7 ± 8.5
Euclidean branch	–	–	✓	51.3 ± 4.6	50.9 ± 5.2	50.4 ± 5.3	48.7 ± 6.1	41.4 ± 6.9	31.0 ± 5.9	19.9 ± 7.5
Geodesic branch	–	–	✓	50.4 ± 4.2	49.6 ± 5.4	50.4 ± 4.0	48.6 ± 6.0	42.2 ± 6.0	33.5 ± 5.8	23.6 ± 6.1
Pseudo-label [99]	✓	–	–	50.1 ± 4.6	50.5 ± 5.5	50.2 ± 4.9	49.3 ± 5.4	46.1 ± 6.2	35.8 ± 6.9	27.9 ± 6.8
	–	✓	–	50.1 ± 4.6	50.7 ± 5.1	50.6 ± 5.1	48.1 ± 4.7	45.2 ± 6.2	34.3 ± 6.2	27.9 ± 6.0
	–	–	✓	50.1 ± 4.6	50.9 ± 5.0	50.1 ± 5.0	48.5 ± 5.1	45.8 ± 5.1	35.1 ± 6.7	26.9 ± 6.3
Con2R [34]	✓	–	–	47.8 ± 6.6	49.6 ± 4.9	49.2 ± 5.6	47.4 ± 6.6	41.4 ± 6.1	30.3 ± 6.7	20.9 ± 6.4
	–	✓	–	47.8 ± 6.6	49.3 ± 5.7	49.1 ± 5.6	47.7 ± 6.3	41.5 ± 6.2	32.3 ± 6.6	21.7 ± 7.2
	–	–	✓	47.8 ± 6.6	48.7 ± 5.2	48.9 ± 6.2	47.7 ± 6.4	42.3 ± 6.6	31.7 ± 5.7	21.5 ± 5.7
FixMatch [102]	–	–	–	52.9 ± 3.9	53.5 ± 4.5	53.6 ± 4.1	53.0 ± 5.1	48.0 ± 6.7	37.8 ± 7.9	22.4 ± 11.7
	✓	–	–	52.9 ± 3.9	53.8 ± 4.5	53.7 ± 4.7	52.4 ± 4.6	49.1 ± 6.9	42.6 ± 8.2	33.0 ± 10.5
	–	✓	–	52.9 ± 3.9	53.3 ± 4.0	52.8 ± 4.7	52.7 ± 4.4	51.2 ± 5.7	43.9 ± 10.5	39.4 ± 8.9
	–	–	✓	52.9 ± 3.9	53.7 ± 4.5	53.5 ± 4.6	52.1 ± 4.7	49.5 ± 6.1	42.5 ± 9.1	32.6 ± 10.8
DSP (Ours)	✓	–	–	53.4 ± 3.6	53.8 ± 5.0	53.5 ± 4.5	52.0 ± 5.6	49.8 ± 6.3	42.4 ± 8.2	32.5 ± 10.4
	–	✓	–	53.4 ± 3.6	53.7 ± 4.8	53.0 ± 3.6	52.9 ± 5.0	50.9 ± 5.7	47.7 ± 7.0	44.2 ± 10.1
	–	–	✓	53.4 ± 3.6	53.7 ± 4.4	53.5 ± 4.9	53.4 ± 4.8	52.3 ± 4.6	47.7 ± 8.2	37.7 ± 12.5

Table A13: Segmentation results of semi-weakly supervised algorithms for cell organelle segmentation on the *HELA-2* dataset in numerical form obtained at increasing *Annotation Compression Ratios* and with different annotation types. Numbers are reported in average mean DICE and standard deviation along 10 cross-validation experiments. Random baseline: 5.1 ± 0.3 DICE. Class-prior baseline: 6.1 ± 0.6 DICE.

In Figure 21 and Figure 22 of chapter 6 the segmentation results are displayed as graphs with the *Annotation Compression Ratio* on the x-axis and the average mean

Method	mixed	$ACR = 1$	$ACR = 2$	$ACR = 4$	$ACR = 8$	$ACR = 16$	$ACR = 32$	$ACR = 64$
				HELA-2				
UNet	–	50.1 ± 4.6	50.3 ± 5.6	48.2 ± 4.9	43.6 ± 7.0	34.2 ± 6.6	24.6 ± 3.7	20.2 ± 6.5
FixMatch [102]	✓	52.9 ± 3.9	$\mathbf{53.8 \pm 4.4}$	53.0 ± 4.2	52.9 ± 4.7	50.2 ± 6.3	46.0 ± 7.4	36.7 ± 11.6
DSP (Ours)	✓	$\mathbf{53.4 \pm 3.6}$	53.7 ± 4.5	$\mathbf{54.0 \pm 4.0}$	$\mathbf{54.2 \pm 4.6}$	$\mathbf{52.1 \pm 5.6}$	$\mathbf{51.6 \pm 6.0}$	$\mathbf{49.5 \pm 6.1}$
				HELA-3				
UNet	–	47.3 ± 7.2	45.6 ± 7.4	44.8 ± 8.4	41.6 ± 8.5	42.0 ± 6.5	34.5 ± 6.4	28.8 ± 9.4
FixMatch [102]	✓	54.2 ± 3.4	$\mathbf{54.5 \pm 3.7}$	$\mathbf{55.9 \pm 2.8}$	$\mathbf{55.7 \pm 2.7}$	54.6 ± 3.0	48.7 ± 4.1	49.4 ± 5.4
DSP (Ours)	✓	$\mathbf{54.5 \pm 3.9}$	53.9 ± 5.0	54.6 ± 3.0	55.5 ± 2.1	$\mathbf{56.2 \pm 1.7}$	$\mathbf{53.3 \pm 5.0}$	$\mathbf{52.4 \pm 4.2}$
				MACROPHAGE-2				
UNet	–	30.7 ± 5.5	29.2 ± 7.7	22.1 ± 4.2	19.0 ± 6.3	10.5 ± 7.5	5.7 ± 2.7	9.3 ± 6.9
FixMatch [102]	✓	44.3 ± 2.1	42.5 ± 7.2	40.1 ± 6.7	23.7 ± 8.4	12.2 ± 6.8	11.7 ± 8.4	13.3 ± 18.8
DSP (Ours)	✓	$\mathbf{45.0 \pm 2.6}$	$\mathbf{43.6 \pm 8.5}$	$\mathbf{43.0 \pm 8.9}$	$\mathbf{31.3 \pm 10.4}$	$\mathbf{21.1 \pm 8.9}$	$\mathbf{14.5 \pm 11.4}$	$\mathbf{15.3 \pm 14.8}$
				JURKAT-1				
UNet	–	28.4 ± 8.2	26.4 ± 6.3	21.1 ± 5.1	14.2 ± 7.1	5.0 ± 2.0	8.2 ± 3.4	5.1 ± 1.9
FixMatch [102]	✓	32.7 ± 9.6	$\mathbf{32.6 \pm 8.6}$	$\mathbf{32.3 \pm 9.5}$	15.7 ± 7.1	5.7 ± 3.5	6.3 ± 3.9	$\mathbf{6.5 \pm 2.7}$
DSP (Ours)	✓	$\mathbf{32.9 \pm 9.1}$	32.4 ± 8.3	29.5 ± 5.3	$\mathbf{18.0 \pm 7.6}$	$\mathbf{7.8 \pm 4.1}$	$\mathbf{10.5 \pm 5.9}$	6.3 ± 3.6

Table A14: Segmentation results of semi-weakly supervised training of FixMatch and *DSP* algorithms with diverse annotation types (masks, bounding boxes, points, image-level labels, unlabaled data), lower baseline Unet trained with masks. Numbers reported in average mean DICE and standard deviation along 10 cross-validation splits for *HELA-2*, 5 splits for *HELA-3*, *MACROPHAGE-2* and *JURKAT-1*.

DICE on the y-axis. This way of displaying the results makes interpreting the progression of segmentation performance with fewer and fewer pixel-wise annotations more graspable. For completeness, here, we provide the numerical counterpart to these graphs. In Table A13 the numerical results of the tested algorithms on the *HELA-2* dataset are displayed, which include experiments using different pairs of annotation types for training. Numerical results for training with the complete mix of annotation types we present in Table A14, which includes results from all the cell organelle datasets, *i.e.* *HELA-2*, *HELA-3*, *MACROPHAGE-2*, and *JURKAT-1*. In Table A13, we further include the semi-supervised results of FixMatch [102], which can be used to directly compare to the semi-weakly versions which make use of

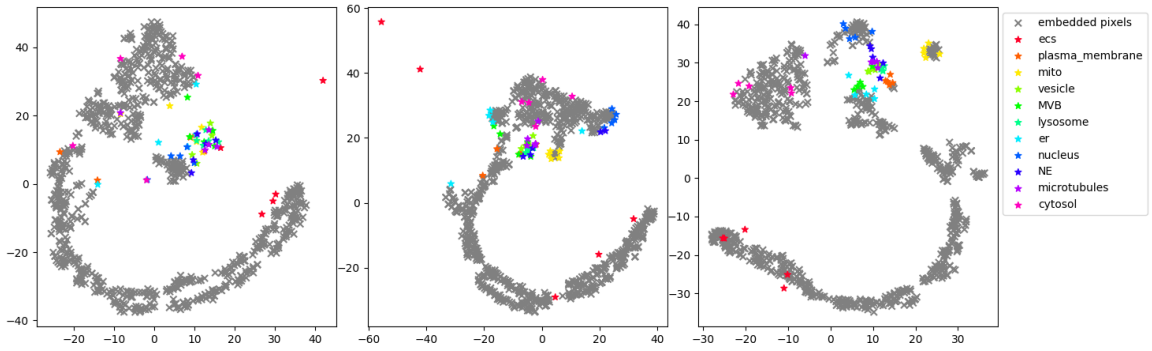


Figure A4: Three plots of the learned class-wise prototypes (colored stars) and randomly selected pixel-embeddings (gray crosses) throughout the training process (left to right: after 10, 50 and 100 epochs of training). Prototypes and embeddings are projected via t-SNE.

pseudo-label filtering. Especially for high ACRs, *i.e.* $16, 32, 64$, we can clearly grasp that the added filtering process makes FixMatch a strong baseline for our training scenarios.

C.3 Additional qualitative insights for Decoupled Semantic Prototypes

In *DSP* we train an altered segmentation architecture where semantic prototypes are used to assign class correspondence. We visualize the learned prototypes as well as randomly selected pixel-embeddings via t-SNE projection [258] in Figure A4. There, we show how the semantic prototypes form clusters which delimit the semantic regions in the embedding space during training after the 10th, 50th and 100th epoch.

C.4 Additional details on the OpenOrganelle dataset, classes and pre-processing

In the dataset description of the OpenOrganelle data [9], we outlined the biologically motivated merging process for which we use the class hierarchy from the code of the

Class name	*HELA-2*	*HELA-3*	*MACROPHAGE-2*	*JURKAT-1*
Extracellular Space	✓	✓	✓	✓
Plasma Membrane	✓	✓	✓	
Mitochondria	✓	✓		✓
Vesicle	✓	✓	✓	✓
Multivesicular bodies	✓	✓	✓	✓
Lysosome	✓		✓	
Endoplasmic Reticulum	✓	✓	✓	✓
Nucleus	✓	✓	✓	✓
Nucelear Envelope	✓	✓		
Microtubule	✓	✓		✓
Cytosol	✓	✓	✓	✓

Figure A5: Classes from the OpenOrganelle dataset [9] which satisfy the cross-validation requirements of our evaluation protocol.

original publication[1]. For the different datasets of OpenOrganelle, a varying amount of classes satisfy the cross-validation requirement of being present in at least three sub-volumes, which we show in Figure A5.
Next, we show the average number of images within each train-, validation and test split of the different datasets:

- *HELA-2*: 2321 training images, 924 validation images, and 930 testing images
- *HELA-3*: 1634 training images, 731 validation images, and 791 testing images
- *MACROPHAGE-2*: 1482 training images, 685 validation images, and 740 testing images
- *JURKAT-1*: 1525 training images, 745 validation images, and 742 testing images

A central part in our experiments is processing image data with a wide diversity of annotation types. To achieve this, we took the pixel-wise annotations of the OpenOrganelle dataset and derived weak annotations from them. Creating image-level labels

[1] https://github.com/saalfeldlab/CNNectome/blob/7c5250edf2ba8ce43127c457b755ea30721f638f/CNNectome/utils/hierarchy.py

merely includes counting the unique classes within the mask, while bounding boxes includes computing the connected components of the pixel-wise mask and drawing a box around each of these components, while concurrently saving the class membership of the component. To set up point annotations, there are several valid choices. We opted for modeling point clicks by taking inspiration from psychology which states that humans generally point at objects by clicking on the medial axis [259] or the center of regions [202]. Therefore, we computed the medoids for each connected component of the pixel-wise annotation and saving its location and the class there.

C.5 Training details and baseline descriptions

In constructing the batches for training semi-weakly supervised learning methods, we make sure that each sampled mini-batch consists on average of all annotation types to equal portions. This can be achieved by over-sampling images which are annotated with a less frequent annotation type, which is a strategy common to semi-supervised learning where the small portion of mask annotated images are heavily over-sampled [102, 34].

Pseudo-label [99]: We make use of online pseudo-labeling, where we compute the pseudo-labels on the fly while training.

FixMatch [102]: To adapt FixMatch to the semi-weakly supervised segmentation setting, in Table A15, we investigated the effect of different thresholds as proposed in the original publication. There, we found that not applying a threshold yields the best results. In the same table, we also investigated which strong augmentations lead to the best segmentation results. For this, we analyzed the effect of applying CutOut [240] with a size of 32×32 multiple times at random locations of the input image with a probability of 50%. The results in Table A15 indicate, that the best results were achieved when applying it up to nine times. This strong augmentation is used for FixMatch, Con2R and our *DSP* method.

threshold	DICE
0.0	51.7 ± 3.6
0.1	51.3 ± 4.2
0.2	51.2 ± 3.9
0.3	51.4 ± 3.5
0.4	51.4 ± 4.2
0.5	51.0 ± 4.0
0.6	51.6 ± 3.9
0.7	51.0 ± 3.7
0.8	51.0 ± 3.6
0.9	50.8 ± 3.6
0.95	50.5 ± 3.4

# CutOut	DICE
0	48.5 ± 3.7
1	49.0 ± 4.1
2	50.0 ± 3.1
3	50.3 ± 3.6
4	50.4 ± 4.0
5	50.4 ± 4.0
6	50.4 ± 4.0
7	50.8 ± 3.5
8	50.8 ± 4.2
9	51.7 ± 3.6
10	51.1 ± 3.4

Table A15: Experiments for FixMatch [102] reported in average mDICE on the *HELA-2* dataset and ACR = 8 scenario. Left: Ablation of the pseudo-label threshold. Right: Ablation of the maximum amount of CutOuts in strong augmentations.

Classification branch [186]: For this baseline, we follow the description of Mlynarski *et al.* [186] as close as possible in order to augment the Unet backbone segmentation network [44] with an additional classification branch. To address the difference in image size as used in the original implementation, we add four more convolutions with ReLU activations after the mean pooling operation and a single convolution of size 11×11 to end up at the matching size as in the paper. Thus, after this, we apply the exact classification branch with linear layers, ReLU and residual connections as Mlynarski *et al.*. To make sure we have the strongest possible variant of this method, we investigated whether the idea of Bae *et al.* [185], *i.e.* using the classification prediction to constrain the segmentation output in inference can help in producing more accurate results. For our use-case of cell organelle segmentation, this process did not help. It is most likely only successful if the classification predictions are exceptional.

Euclidean/Geodesic point branch: We designed baselines for the scenario of training with pixel-wise mask annotations as well as point annotations. For this we train multi-task Unets [44] which have an additional output-head to regress distance maps which are generated from the point annotations. Therefore, for all points of the same class within an image we compute a class-specific distance map, where each value in this map indicates the smallest distance to one of the given points.

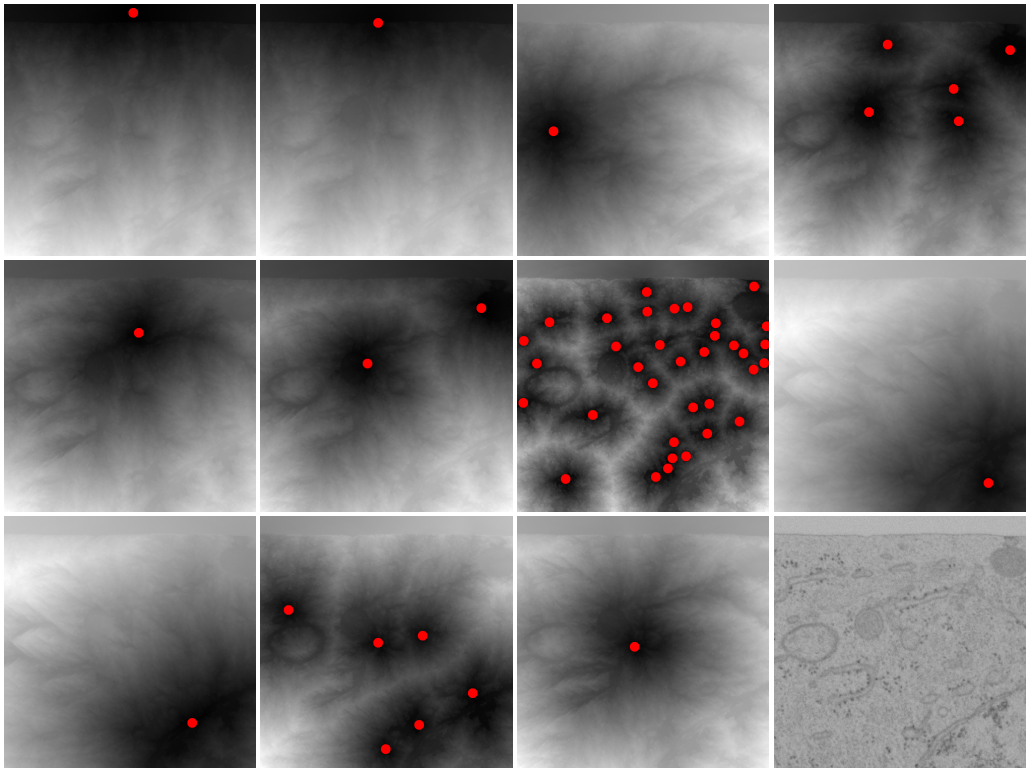


Figure A6: Class-specific geodesic distances for point cues (red). The leftmost image in the bottom row is the input image from the *HELA-2* dataset.

In Figure A6, we visualize distance maps based on the geodesic distance for an image with point annotations, which indicates the rich structural information the model learns through regressing these maps during training. We choose the geodesic- and the euclidean distance for the compuation of these distance maps. The choice of the geodesic distance is motivated by the integration of click cues from interactive segmentation [252] and weakly supervised medical segmentation [260]. Integrating a loss based on the regression of distance maps is commonly explored in the medical fieled as well as more specifically for cell datasets [261, 262].

Box loss [248]: By incorporating the bounding box-based loss proposed by Tian *et al.* [248], we extend its application to our scenario where both masks and bounding boxes are available. This loss is directly applied to the segmentation output-head.

D Curriculum vitae – Simon Michael Reiß

Education

Oct '17 - Feb '20	**Master's degree in computer science** Karlsruhe Institute of Technology Specialization: study profile artificial intelligence
Oct '13 - Nov '16	**Bachelor's degree in software engineering** University of Stuttgart
Jun '13	**Abitur diploma** Gymnasium in der Taus, Backnang

Academic work experience

Apr '20 - Jul '23	**Research associate** at Computer Vision for Human-Computer Interaction Lab working in a collaboration project with Carl Zeiss AG Topic: *Semantic image segmentation with few and coarse annotations*

Feb '20	**Master's thesis** at Computer Vision for Human-Computer Interaction Lab Supervision: M. Sc. A. Roitberg Examination: Prof. Dr.-Ing. Rainer Stiefelhagen Topic: *Zero-shot recognition of composite activities in context of driver observation*
Aug '18 - Feb '20	**Student research assistant** at Computer Vision for Human-Computer Interaction Lab working on in-vehicle human activity recognition Supervision: M. Sc. A. Roitberg Topics: *Action recognition, image-to-image translation*
Nov '16	**Bachelor's thesis** at Fraunhofer Institute for Industrial Engineering IAO Supervision: M. Sc. Julien Ostermann, M. Sc. Kristian Lehmann, Dipl.-Inf. Sebastian Wagner Examination: Prof. Dr. Dr. h. c. Frank Leymann Topic: *Services for data platforms – module to analyze and process urban time-based sensor data*

Additional experience

Teaching	Supervision of master's theses, supervision of students in the *Practical Course Computer Vision for Human-Computer Interaction* as well as in the *Seminar Computer Vision for Human-Computer Interaction* Preparing and giving lectures for the courses *Deep Learning for Computer Vision I: Basics* and *Deep Learning for Computer Vision II: Advanced Topics*

Reviewing duties	Serving as reviewer for computer vision-related venues such as CVPRW20, IROS21, BMVC21, GCPR21, IV21, CVPR21, ICCV21, WACV22, CVPR22, ECCV22, CVPR23, ICCV23
Conference visits	IEEE Intelligent Vehicles Symposium, 2020, virtual IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, virtual IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, virtual European Conference on Computer Vision, 2022, Tel Aviv, Israel IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, Vancouver, Canada
Award	Participant in the CVPR 2023 Doctoral Consortium

E Authored publications in order of appearance

This doctoral research resulted in the following thesis-related publications:

1. **Every annotation counts: Multi-label deep supervision for medical image segmentation**
   Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, Rainer Stiefelhagen
   Conference on Computer Vision and Pattern Recognition (CVPR), 2021

2. **Graph-Constrained Contrastive Regularization for Semi-weakly Volumetric Segmentation**
   Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, Rainer Stiefelhagen
   European Conference on Computer Vision (ECCV), 2022

3. **Decoupled Semantic Prototypes enable learning from diverse annotation types for semi-weakly segmentation in expert-driven domains**
   Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, Rainer Stiefelhagen
   Conference on Computer Vision and Pattern Recognition (CVPR), 2023

The following publications were co-authored by Simon Reiß but are not directly related to the thesis content:

1. **Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles**
   Manuel Martin, Alina Roitberg, Monica Haurilet, Matthias Horne, Simon Reiß, Michael Voit, Rainer Stiefelhagen
   International Conference on Computer Vision (ICCV), 2019

2. **Activity-aware attributes for zero-shot driver behavior recognition**
   Simon Reiß, Alina Roitberg, Monica Haurilet, Rainer Stiefelhagen

Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020

3. **Cnn-based driver activity understanding: Shedding light on deep spatiotemporal representations**
   Alina Roitberg, Monica Haurilet, Simon Reiß, Rainer Stiefelhagen
   International Conference on Intelligent Transportation Systems (ITS), 2020

4. **Deep classification-driven domain adaptation for cross-modal driver behavior recognition**
   Simon Reiß, Alina Roitberg, Monica Haurilet, Rainer Stiefelhagen
   Intelligent Vehicles Symposium (IV), 2020

5. **Capturing omni-range context for omnidirectional segmentation**
   Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, Rainer Stiefelhagen
   Conference on Computer Vision and Pattern Recognition (CVPR), 2021

6. **From Driver Talk To Future Action: Vehicle Maneuver Prediction by Learning from Driving Exam Dialogs**
   Alina Roitberg, Simon Reiß, Rainer Stiefelhagen
   Intelligent Vehicles Symposium (IV), 2021

7. **Let's play for action: Recognizing activities of daily living by learning from life simulation video games**
   Alina Roitberg, David Schneider, Aulia Djamal, Constantin Seibold, Simon Reiß, Rainer Stiefelhagen
   International Conference on Intelligent Robots and Systems (IROS), 2021

8. **Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation**
   Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, Rainer Stiefelhagen
   Conference on Computer Vision and Pattern Recognition (CVPR), 2022

9. **Reference-guided pseudo-label generation for medical semantic segmentation**
   Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, Rainer Stiefelhagen
   AAAI Conference on Artificial Intelligence (AAAI), 2022

10. **Breaking with fixed set pathology recognition through report-guided contrastive training**
    Constantin Seibold, Simon Reiß, M Saquib Sarfraz, Rainer Stiefelhagen, Jens Kleesiek
    Medical Image Computing and Computer Assisted Intervention (MICCAI), 2022

11. **Detailed Annotations of Chest X-Rays via CT Projection for Report Understanding**
    Constantin Seibold, Simon Reiß, Saquib Sarfraz, Matthias A Fink, Victoria Mayer, Jan Sellner, Moon Sung Kim, Klaus H Maier-Hein, Jens Kleesiek, Rainer Stiefelhagen
    British Machine Vision Conference (BMVC), 2022

12. **Delivering Arbitrary-Modal Semantic Segmentation**
    Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, Rainer Stiefelhagen
    Conference on Computer Vision and Pattern Recognition (CVPR), 2023

Bibliography

[1] Michael Chui, James Manyika, Mehdi Miremadi, Nicolaus Henke, Rita Chung, Pieter Nel, and Sankalp Malhotra. Notes from the ai frontier: Applications and value of deep learning.

[2] Science News Staff. Ai is changing how we do science. get a glimpse.

[3] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[6] Andrew Ng. How to unlock ai value.

[7] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M

Summers, Bram van Ginneken, et al. The medical segmentation decathlon. *arXiv preprint arXiv:2106.05735*, 2021.

[8] Hrvoje Bogunović, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, et al. Retouch: the retinal oct fluid detection and segmentation benchmark and challenge. *IEEE transactions on medical imaging*, 38(8):1858–1874, 2019.

[9] Larissa Heinrich, Davis Bennett, David Ackerman, Woohyun Park, John Bogovic, Nils Eckstein, Alyson Petruncio, Jody Clements, Song Pang, C Shan Xu, et al. Whole-cell organelle segmentation in volume electron microscopy. *Nature*, 599(7883):141–146, 2021.

[10] 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. *arXiv preprint arXiv:2304.02643*, 2023.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

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

[14] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[15] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

[16] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[17] Yaqi Wang, Shuai Wang, Fan Ye, Weiwei Cui, Yifan Zhang, Liaoyuan Zeng, and Xingru Huang. Semi-supervised teeth segmentation. Apr 2023.

[18] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[19] Spyridon Bakas, Ujjwal Baid, Keyvan Farahani, Jake Albrecht, James Eddy, Timothy Bergquist, Thomas Yu, Verena Chung, Russell (Taki) Shinohara, Michel Bilello, and et al. The international brain tumor segmentation (brats) cluster of challenges. Apr 2023.

[20] Jun Ma and Bo Wang. Fast, low-resource, and accurate organ and pan-cancer segmentation in abdomen ct. Apr 2023.

[21] Kuestner T. Gatidis S. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions (fdg-pet-ct-lesions) [dataset]. *The Cancer Imaging Archive*, 2022.

[22] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, Yoel Shoshan, Flora Gilboa-Solomon, Yasmeen George, Xi Yang, Jianpeng Zhang, Jing Zhang, Yong Xia, Mengran Wu, Zhiyang Liu, Ed Walczak, Sean McSweeney, Ranveer Vasdev, Chris Hornung, Rafat Solaiman, Jamee Schoephoerster, Bailey Abernathy, David Wu, Safa Abdulkadir, Ben Byun, Justice Spriggs, Griffin Struyk, Alexandra Austin, Ben Simpson, Michael Hagstrom, Sierra Virnig, John French, Nitin Venkatesh, Sarah Chan, Keenan Moore, Anna Jacobsen, Susan Austin, Mark Austin, Subodh Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct, 2023.

[23] Kuanquan Wang, Mingwang Xu, Qiucheng Wang, Wen Cheng, Wei Wang, and Xinjie Liang. Tumor detection, segmentation, and classification challenge on automated 3d breast ultrasound. Mar 2022.

[24] Xiangde Luo, Wenjun Liao, Mu Zhou, Jia Fu, Shichuan Zhang, Guotai Wang, and Shaoting Zhang. Segmentation of organs-at-risk and gross tumor volume for radiotherapy planning of nasopharyngeal carcinoma challenge 2023. Apr 2023.

[25] Patrick Carnahan, Apurva Bharucha, Mehdi Eskandari, Elvis C.S. Chen, and Terry M. Peters. Segmentation of the mitral valve from 3d transesophageal echocardiography. Apr 2023.

[26] Rina Bao, Yangming Ou, and P. Ellen Grant. Hypoxic ischemic encephalopathy lesion segmentation challenge. Apr 2023.

[27] Kaiyuan Yang, Hongwei Bran Li, Anjany Sekuboyina, Bjoern Menze, Susanne Wegener, Yihui Ma, Laura Westphal, Rami Al-Maskari, Luciano Höher, Fabio

Musio, and et al. Topology-aware anatomical segmentation of the circle of willis for cta and mra. Apr 2023.

[28] Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, et al. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, 2023.

[29] Marco Cipriano, Stefano Allegretti, Federico Bolelli, Mattia Di Bartolomeo, Federico Pollastri, Arrigo Pellacani, Paolo Minafra, Alexandre Anesi, and Costantino Grana. Deep segmentation of the mandibular canal: a new 3d annotated dataset of cbct volumes. *IEEE Access*, 10:11500–11510, 2022.

[30] Félix Quinton, Romain Popoff, Benoît Presles, Sarah Leclerc, Fabrice Meriaudeau, Guillaume Nodari, Olivier Lopez, Julie Pellegrinelli, Olivier Chevallier, Dominique Ginhac, et al. A tumour and liver automatic segmentation (atlas) dataset on contrast-enhanced magnetic resonance imaging for hepatocellular carcinoma. *Data*, 8(5):79, 2023.

[31] The Medical Image Computing and Computer Assisted Intervention Society. Miccai registered challenges.

[32] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. Preparing medical imaging data for machine learning. *Radiology*, 295(1):4–15, 2020.

[33] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Every annotation counts: Multi-label deep supervision for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9532–9542, 2021.

[34] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Graph-constrained contrastive regularization for semi-weakly volumetric segmentation. In *European Conference on Computer Vision*, pages 401–419. Springer, 2022.

[35] Simon Reiß, Constantin Seibold, Alexander Freytag, Erik Rodner, and Rainer Stiefelhagen. Decoupled semantic prototypes enable learning from diverse annotation types for semi-weakly segmentation in expert-driven domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15495–15506, June 2023.

[36] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[37] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[38] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127:302–321, 2019.

[39] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[40] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.

[41] 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*, pages 801–818, 2018.

[42] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017.

[43] 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*, pages 3431–3440, 2015.

[44] 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*, pages 234–241. Springer, 2015.

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017.

[46] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[47] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[48] 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*, pages 770–778, 2016.

[49] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[50] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015.

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

[52] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11534–11542, 2020.

[53] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. *arXiv preprint arXiv:1805.10180*, 2018.

[54] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7151–7160, 2018.

[55] Yuhui Yuan, Lang Huang, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv preprint arXiv:1809.00916*, 2018.

[56] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3146–3154, 2019.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[58] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[59] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[60] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[61] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[62] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882*, 2020.

[63] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021.

[64] 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*, pages 248–255. Ieee, 2009.

[65] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[66] 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*, pages 424–432. Springer, 2016.

[67] 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*, pages 565–571. IEEE, 2016.

[68] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 3–11. Springer, 2018.

[69] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.

[70] Xin Yang, Lequan Yu, Shengli Li, Xu Wang, Na Wang, Jing Qin, Dong Ni, and Pheng-Ann Heng. Towards automatic semantic segmentation in volumetric ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 711–719. Springer, 2017.

[71] Andriy Myronenko. 3d mri brain tumor segmentation using autoencoder regularization. In *International MICCAI Brainlesion Workshop*, pages 311–320. Springer, 2018.

[72] Ali Hatamizadeh, Demetri Terzopoulos, and Andriy Myronenko. Edge-gated cnns for volumetric semantic segmentation of medical images. *arXiv preprint arXiv:2002.04207*, 2020.

[73] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. No new-net. In *International MICCAI Brainlesion Workshop*, pages 234–244. Springer, 2018.

[74] Fabian Isensee and Klaus H Maier-Hein. An attempt at beating the 3d u-net. *arXiv preprint arXiv:1908.02182*, 2019.

[75] Richard McKinley, Raphael Meier, and Roland Wiest. Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 456–465. Springer, 2018.

[76] Guangrui Mu, Zhiyong Lin, Miaofei Han, Guang Yao, and Yaozong Gao. Segmentation of kidney tumor by multi-resolution vb-nets. 2019.

[77] Yao Zhang, Yixin Wang, Feng Hou, Jiawei Yang, Guangwei Xiong, Jiang Tian, and Cheng Zhong. Cascaded volumetric convolutional network for kidney tumor segmentation from ct volumes. *arXiv preprint arXiv:1910.02235*, 2019.

[78] Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The kits21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase ct. *arXiv preprint arXiv:2307.01984*, 2023.

[79] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong

encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*, 2021.

[80] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[81] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, pages 205–218. Springer, 2022.

[82] 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). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.

[83] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

[84] Jens Kleesiek, Gregor Urban, Alexander Hubert, Daniel Schwarz, Klaus Maier-Hein, Martin Bendszus, and Armin Biller. Deep mri brain extraction: A 3d convolutional neural network for skull stripping. *NeuroImage*, 129:460–469, 2016.

[85] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Deep end2end voxel2voxel prediction. In *Proceedings of the IEEE*

conference on computer vision and pattern recognition workshops, pages 17–24, 2016.

[86] Sergios Gatidis, Marcel Früh, Matthias Fabritius, Sijing Gu, Konstantin Nikolaou, Christian La Fougère, Jin Ye, Junjun He, Yige Peng, Lei Bi, et al. The autopet challenge: Towards fully automated lesion segmentation in oncologic pet/ct imaging. 2023.

[87] Nabila Abraham and Naimul Mefraz Khan. A novel focal tversky loss function with improved attention u-net for lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging*, pages 683–687. IEEE, 2019.

[88] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE transactions on medical imaging*, 37(7):1597–1605, 2018.

[89] Bingbing Li, Chengdong Wu, Jianning Chi, Xiaosheng Yu, and Gang Wang. A deeply supervised convolutional neural network for brain tumor segmentation. In *2020 39th Chinese Control Conference (CCC)*, pages 6262–6267. IEEE, 2020.

[90] Yan Xu, Yang Li, Mingyuan Liu, Yipei Wang, Maode Lai, I Eric, and Chao Chang. Gland instance segmentation by deep multichannel side supervision. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 496–504. Springer, 2016.

[91] Yishuo Zhang and Albert CS Chung. Deep supervision with additional labels for retinal vessel segmentation task. In *International conference on medical image computing and computer-assisted intervention*, pages 83–91. Springer, 2018.

[92] Qi Dou, Lequan Yu, Hao Chen, Yueming Jin, Xin Yang, Jing Qin, and Pheng-Ann Heng. 3d deeply supervised network for automated segmentation of volumetric medical images. *Medical image analysis*, 41:40–54, 2017.

[93] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Artificial intelligence and statistics*, pages 562–570, 2015.

[94] Kaiqiang Chen, Michael Weinmann, Xian Sun, Menglong Yan, Stefan Hinz, Boris Jutzi, and Martin Weinmann. Semantic segmentation of aerial imagery via multi-scale shuffling convolutional neural networks with deep supervision. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 4(1), 2018.

[95] Yun Liu, Ming-Ming Cheng, Xinyu Zhang, Guang-Yu Nie, and Meng Wang. Dna: Deeply-supervised nonlinear aggregation for salient object detection. *arXiv preprint arXiv:1903.12476*, 2019.

[96] Dimitrios Marmanis, Jan D Wegner, Silvano Galliani, Konrad Schindler, Mihai Datcu, and Uwe Stilla. Semantic segmentation of aerial images with an ensemble of cnss. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016*, 3:473–480, 2016.

[97] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[98] Yuyin Zhou, Lingxi Xie, Elliot K Fishman, and Alan L Yuille. Deep supervision for pancreatic cyst segmentation in abdominal ct scans. In *International conference on medical image computing and computer-assisted intervention*, pages 222–230. Springer, 2017.

[99] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013.

[100] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4268–4277, June 2022.

[101] 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*, pages 253–260. Springer, 2017.

[102] Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *arXiv preprint arXiv:2001.07685*, 2020.

[103] Constantin Marc Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2171–2179, 2022.

[104] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in neural information processing systems*, pages 1195–1204, 2017.

[105] Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, pages 12674–12684, 2020.

[106] Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4258–4267, 2022.

[107] Christian S Perone and Julien Cohen-Adad. Deep semi-supervised segmentation with weight-averaged consistency targets. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, pages 12–19. Springer, 2018.

[108] Geoff French, Timo Aila, Samuli Laine, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, high-dimensional perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

[109] Viktor Olsson, Wilhelm Tranheden, Juliano Pinto, and Lennart Svensson. Classmix: Segmentation-based data augmentation for semi-supervised learning. *arXiv preprint arXiv:2007.07936*, 2020.

[110] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6023–6032, 2019.

[111] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021.

[112] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a“ siamese” time delay neural network. *Advances in neural information processing systems*, 6, 1993.

[113] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[114] Chun-Hsiao Yeh, Cheng-Yao Hong, Yen-Chi Hsu, Tyng-Luh Liu, Yubei Chen, and Yann LeCun. Decoupled contrastive learning. In *European Conference on Computer Vision*, pages 668–684. Springer, 2022.

[115] Yuanyi Zhong, Bodi Yuan, Hong Wu, Zhiqiang Yuan, Jian Peng, and Yu-Xiong Wang. Pixel contrastive-consistent semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7273–7282, 2021.

[116] Xiangyun Zhao, Raviteja Vemulapalli, Philip Andrew Mansfield, Boqing Gong, Bradley Green, Lior Shapira, and Ying Wu. Contrastive learning for label efficient semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10623–10633, 2021.

[117] Zhenda Xie, Yutong Lin, Zheng Zhang, Yue Cao, Stephen Lin, and Han Hu. Propagate yourself: Exploring pixel-level consistency for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16684–16693, 2021.

[118] Inigo Alonso, Alberto Sabater, David Ferstl, Luis Montesano, and Ana C Murillo. Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8219–8228, 2021.

[119] Yanning Zhou, Hang Xu, Wei Zhang, Bin Gao, and Pheng-Ann Heng. C3-semiseg: Contrastive semi-supervised segmentation via cross-set learning and dynamic class-balancing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7036–7045, 2021.

[120] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[121] Dong Nie, Yaozong Gao, Li Wang, and Dinggang Shen. Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 370–378. Springer, 2018.

[122] 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)*, pages 121–140. IEEE, 2019.

[123] 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*, pages 3646–3655, 2020.

[124] Yingda Xia, Dong Yang, Zhiding Yu, Fengze Liu, Jinzheng Cai, Lequan Yu, Zhuotun Zhu, Daguang Xu, Alan Yuille, and Holger Roth. Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. *Medical Image Analysis*, 65:101766, 2020.

[125] Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 605–613. Springer, 2019.

[126] Xiangde Luo, Wenjun Liao, Jieneng Chen, Tao Song, Yinan Chen, Shichuan Zhang, Nianyong Chen, Guotai Wang, and Shaoting Zhang. Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 318–329. Springer, 2021.

[127] Chenyu You, Ruihan Zhao, Lawrence Staib, and James S Duncan. Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *arXiv preprint arXiv:2105.07059*, 2021.

[128] Chenyu You, Yuan Zhou, Ruihan Zhao, Lawrence Staib, and James S Duncan. Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. *arXiv preprint arXiv:2108.06227*, 2021.

[129] Dewen Zeng, Yawen Wu, Xinrong Hu, Xiaowei Xu, Haiyun Yuan, Meiping Huang, Jian Zhuang, Jingtong Hu, and Yiyu Shi. Positional contrastive learning for volumetric medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 221–230. Springer, 2021.

[130] Yassine Ouali, Céline Hudelot, and Myriam Tami. An overview of deep semi-supervised learning. *arXiv preprint arXiv:2006.05278*, 2020.

[131] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[132] Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5070–5079, 2019.

[133] Junnan Li, Caiming Xiong, and Steven CH Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9475–9484, 2021.

[134] Siddhartha Chandra and Iasonas Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *European conference on computer vision*, pages 402–418. Springer, 2016.

[135] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.

[136] Philipp Krähenbühl and Vladlen Koltun. Parameter learning and convergent inference for dense random fields. In *International Conference on Machine Learning*, pages 513–521. PMLR, 2013.

[137] Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip HS Torr. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1529–1537, 2015.

[138] Deepak Pathak, Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional multi-class multiple instance learning. *arXiv preprint arXiv:1412.7144*, 2014.

[139] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8991–9000, 2020.

[140] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7223–7233, 2019.

[141] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7014–7023, 2018.

[142] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016.

[143] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5267–5276, 2019.

[144] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018.

[145] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020.

[146] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016.

[147] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1818–1827, 2018.

[148] Bin Wang, Guojun Qi, Sheng Tang, Tianzhu Zhang, Yunchao Wei, Linghui Li, and Yongdong Zhang. Boundary perception guidance: A scribble-supervised semantic segmentation approach. In *IJCAI International joint conference on artificial intelligence*, 2019.

[149] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 507–522, 2018.

[150] Amy Bearman, Olga Russakovsky, Vittorio Ferrari, and Li Fei-Fei. What's the point: Semantic segmentation with point supervision. In *European conference on computer vision*, pages 549–565. Springer, 2016.

[151] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala. Material recognition in the wild with the materials in context database. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3479–3487, 2015.

[152] Hongjun Chen, Jinbao Wang, Hong Cai Chen, Xiantong Zhen, Feng Zheng, Rongrong Ji, and Ling Shao. Seminar learning for click-level weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6920–6929, 2021.

[153] Yanwei Li, Hengshuang Zhao, Xiaojuan Qi, Yukang Chen, Lu Qi, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Fully convolutional networks for panoptic segmentation with point-based supervision. *arXiv preprint arXiv:2108.07682*, 2021.

[154] Viveka Kulharia, Siddhartha Chandra, Amit Agrawal, Philip Torr, and Ambrish Tyagi. Box2seg: Attention weighted loss and discriminative feature learning for weakly supervised segmentation. In *European Conference on Computer Vision*, pages 290–308. Springer, 2020.

[155] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3136–3145, 2019.

[156] Qizhu Li, Anurag Arnab, and Philip HS Torr. Weakly-and semi-supervised panoptic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 102–118, 2018.

[157] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 876–885, 2017.

[158] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1635–1643, 2015.

[159] Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10, 1997.

[160] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[161] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. 2014.

[162] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[163] Thomas Schlegl, Sebastian M Waldstein, Wolf-Dieter Vogl, Ursula Schmidt-Erfurth, and Georg Langs. Predicting semantic descriptions from medical images with convolutional neural networks. In *International Conference on Information Processing in Medical Imaging*, pages 437–448. Springer, 2015.

[164] Hyun-Lim Yang, Jong Jin Kim, Jong Ho Kim, Yong Koo Kang, Dong Ho Park, Han Sang Park, Hong Kyun Kim, and Min-Soo Kim. Weakly supervised lesion localization for age-related macular degeneration detection using optical coherence tomography images. *PloS one*, 14(4):e0215076, 2019.

[165] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.

[166] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004.

[167] Pablo Arbeláez, Jordi Pont-Tuset, Jonathan T Barron, Ferran Marques, and Jitendra Malik. Multiscale combinatorial grouping. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 328–335, 2014.

[168] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[169] Saeedeh Afshari, Aïcha BenTaieb, Zahra Mirikharaji, and Ghassan Hamarneh. Weakly supervised fully convolutional network for pet lesion segmentation. In *Medical Imaging 2019: Image Processing*, volume 10949, page 109491K. International Society for Optics and Photonics, 2019.

[170] Hoel Kervadec, Jose Dolz, Shanshan Wang, Eric Granger, and Ismail Ben Ayed. Bounding boxes for weakly supervised segmentation: Global constraints get close to full supervision. *arXiv preprint arXiv:2004.06816*, 2020.

[171] 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. *IEEE transactions on medical imaging*, 36(2):674–683, 2016.

[172] Hyeonsoo Lee and Won-Ki Jeong. Scribble2label: Scribble-supervised cell segmentation via self-generating pseudo-labels with consistency. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 14–23. Springer, 2020.

[173] Ke Zhang and Xiahai Zhuang. Cyclemix: A holistic strategy for medical image segmentation from scribble supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11656–11665, 2022.

[174] Gabriele Valvano, Andrea Leo, and Sotirios A Tsaftaris. Learning to segment from scribbles using multi-scale adversarial attention gates. *IEEE Transactions on Medical Imaging*, 40(8):1990–2001, 2021.

[175] Hui Qu, Pengxiang Wu, Qiaoying Huang, Jingru Yi, Zhennan Yan, Kang Li, Gregory M Riedlinger, Subhajyoti De, Shaoting Zhang, and Dimitris N Metaxas. Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. *IEEE transactions on medical imaging*, 39(11):3655–3666, 2020.

[176] Holger Roth, Ling Zhang, Dong Yang, Fausto Milletari, Ziyue Xu, Xiaosong Wang, and Daguang Xu. Weakly supervised segmentation from extreme points. In *Large-Scale Annotation of Biomedical Data and Expert Label Synthesis and Hardware Aware Learning for Medical Imaging and Computer Assisted Intervention*, pages 42–50. Springer, 2019.

[177] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3133–3142, 2020.

[178] Feng Gao, Minhao Hu, Min-Er Zhong, Shixiang Feng, Xuwei Tian, Xiaochun Meng, Zeping Huang, Minyi Lv, Tao Song, Xiaofan Zhang, et al. Segmentation only uses sparse annotations: Unified weakly and semi-supervised learning in medical images. *Medical Image Analysis*, page 102515, 2022.

[179] Pei Wang, Zhaowei Cai, Hao Yang, Gurumurthy Swaminathan, Nuno Vasconcelos, Bernt Schiele, and Stefano Soatto. Omni-detr: Omni-supervised object detection with transformers. *arXiv preprint arXiv:2203.16089*, 2022.

[180] Zhongzheng Ren, Zhiding Yu, Xiaodong Yang, Ming-Yu Liu, Alexander G Schwing, and Jan Kautz. Ufo^2: A unified framework towards omni-supervised object detection. In *European Conference on Computer Vision*, pages 288–313. Springer, 2020.

[181] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in neural information processing systems*, pages 1495–1503, 2015.

[182] Sudhanshu Mittal, Maxim Tatarchenko, and Thomas Brox. Semi-supervised semantic segmentation with high-and low-level consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[183] Nasim Souly, Concetto Spampinato, and Mubarak Shah. Semi supervised semantic segmentation using generative adversarial network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5688–5696, 2017.

[184] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly-and semi-supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7268–7277, 2018.

[185] Wonho Bae, Junhyug Noh, Milad Jalali Asadabadi, and Danica J Sutherland. One weird trick to improve your semi-weakly supervised semantic segmentation model. *arXiv preprint arXiv:2205.01233*, 2022.

[186] Pawel Mlynarski, Hervé Delingette, Antonio Criminisi, and Nicholas Ayache. Deep learning with mixed supervision for brain tumor segmentation. *Journal of Medical Imaging*, 6(3):034002, 2019.

[187] Wenfeng Luo and Meng Yang. Semi-supervised semantic segmentation via strong-weak dual-branch network. In *European Conference on Computer Vision*, pages 784–800. Springer, 2020.

[188] Liyan Sun, Jianxiong Wu, Xinghao Ding, Yue Huang, Zhong Chen, Guisheng Wang, and Yizhou Yu. A teacher-student framework for liver and tumor segmentation under mixed supervision from abdominal ct scans. *Neural Computing and Applications*, pages 1–15, 2022.

[189] Rosana El Jurdi, Caroline Petitjean, Paul Honeine, and Fahed Abdallah. Bb-unet: U-net with bounding box prior. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1189–1198, 2020.

[190] Mostafa S Ibrahim, Arash Vahdat, Mani Ranjbar, and William G Macready. Semi-supervised semantic image segmentation with self-correcting networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2020.

[191] Zhuo Zhao, Lin Yang, Hao Zheng, Ian H Guldner, Siyuan Zhang, and Danny Z Chen. Deep learning based instance segmentation in 3d biomedical images using weak annotation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 352–360. Springer, 2018.

[192] Jose Dolz, Christian Desrosiers, and Ismail Ben Ayed. Teach me to segment with mixed supervision: Confident students become masters. In *International Conference on Information Processing in Medical Imaging*, pages 517–529. Springer, 2021.

[193] Samarth Sinha, Sayna Ebrahimi, and Trevor Darrell. Variational adversarial active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5972–5981, 2019.

[194] Kwanyoung Kim, Dongwon Park, Kwang In Kim, and Se Young Chun. Task-aware variational adversarial active learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8166–8175, 2021.

[195] Shuai Xie, Zunlei Feng, Ying Chen, Songtao Sun, Chao Ma, and Mingli Song. Deal: Difficulty-aware active learning for semantic segmentation. In *Proceedings of the Asian conference on computer vision*, 2020.

[196] S Alireza Golestaneh and Kris M Kitani. Importance of self-consistency in active learning for semantic segmentation. *arXiv preprint arXiv:2008.01860*, 2020.

[197] Sudhanshu Mittal, Maxim Tatarchenko, Özgün Çiçek, and Thomas Brox. Parting with illusions about deep active learning. *arXiv preprint arXiv:1912.05361*, 2019.

[198] Siyu Huang, Tianyang Wang, Haoyi Xiong, Jun Huan, and Dejing Dou. Semi-supervised active learning with temporal output discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3447–3456, 2021.

[199] Zhengfeng Lai, Chao Wang, Luca Cerny Oliveira, Brittany N Dugger, Sen-Ching Cheung, and Chen-Nee Chuah. Joint semi-supervised and active learning for segmentation of gigapixel pathology images with cost-effective labeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 591–600, 2021.

[200] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman. Geodesic star convexity for interactive image segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3129–3136. IEEE, 2010.

[201] 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*, pages 373–381, 2016.

[202] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: Towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022.

[203] Guotai Wang, Wenqi Li, Maria A Zuluaga, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Interactive medical image segmentation using deep learning with image-specific fine tuning. *IEEE transactions on medical imaging*, 37(7):1562–1573, 2018.

[204] Xiangde Luo, Guotai Wang, Tao Song, Jingyang Zhang, Michael Aertsen, Jan Deprest, Sebastien Ourselin, Tom Vercauteren, and Shaoting Zhang. Mideepseg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Medical image analysis*, 72:102102, 2021.

[205] Karol Gotkowski, Camila Gonzalez, Isabel Kaltenborn, Ricarda Fischbach, Andreas Bucher, and Anirban Mukhopadhyay. i3deep: Efficient 3d interactive segmentation with the nnu-net. In *International Conference on Medical Imaging with Deep Learning*, pages 441–456. PMLR, 2022.

[206] Muhammad Asad, Lucas Fidon, and Tom Vercauteren. Econet: Efficient convolutional online likelihood network for scribble-based interactive segmentation. In *International Conference on Medical Imaging with Deep Learning*, pages 35–47. PMLR, 2022.

[207] Basil Mustafa, Aaron Loh, Jan Freyberg, Patricia MacWilliams, Megan Wilson, Scott Mayer McKinney, Marcin Sieniek, Jim Winkens, Yuan Liu, Peggy Bui, et al. Supervised transfer learning at scale for medical imaging. *arXiv preprint arXiv:2101.05913*, 2021.

[208] 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. *Nature*, 577(7788):89–94, 2020.

[209] 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*, volume 33, pages 590–597, 2019.

[210] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram Van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *Medical Image Computing and Computer Assisted Intervention- MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 516–524. Springer, 2017.

[211] Sihong Chen, Kai Ma, and Yefeng Zheng. Med3d: Transfer learning for 3d medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

[212] Wenxuan Wang, Jiachen Shen, Chen Chen, Jianbo Jiao, Yan Zhang, Shanshan Song, and Jiangyun Li. Med-tuning: Exploring parameter-efficient transfer learning for medical volumetric segmentation. *arXiv preprint arXiv:2304.10880*, 2023.

[213] Javier Gamazo Tejero, Martin S. Zinkernagel, Sebastian Wolf, Raphael Sznitman, and Pablo Márquez Neila. Full or weak annotations? an adaptive strategy for budget-constrained annotation campaigns. *arXiv preprint arXiv:2303.11678*, 2023.

[214] Rafid Mahmood, James Lucas, Jose M Alvarez, Sanja Fidler, and Marc T Law. Optimizing data collection for machine learning. *arXiv preprint arXiv:2210.01234*, 2022.

[215] Rafid Mahmood, James Lucas, David Acuna, Daiqing Li, Jonah Philion, Jose M Alvarez, Zhiding Yu, Sanja Fidler, and Marc T Law. How much more data do i need? estimating requirements for downstream tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 275–284, 2022.

[216] Achin Jain, Gurumurthy Swaminathan, Paolo Favaro, Hao Yang, Avinash Ravichandran, Hrayr Harutyunyan, Alessandro Achille, Onkar Dabeer, Bernt Schiele, Ashwin Swaminathan, et al. A meta-learning approach to predicting performance and data requirements. *arXiv preprint arXiv:2303.01598*, 2023.

[217] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[218] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[219] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[220] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[221] David Huang, Eric A Swanson, Charles P Lin, Joel S Schuman, William G Stinson, Warren Chang, Michael R Hee, Thomas Flotte, Kenton Gregory, Carmen A Puliafito, et al. Optical coherence tomography. *science*, 254(5035):1178–1181, 1991.

[222] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[223] Kunihiko Fukushima. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5(4):322–333, 1969.

[224] Benjamin Bischke, Patrick Helber, Joachim Folz, Damian Borth, and Andreas Dengel. Multi-task learning for segmentation of building footprints with deep neural networks. In *2019 IEEE International Conference on Image Processing*, pages 1480–1484. IEEE, 2019.

[225] Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, and Raquel Urtasun. Multinet: Real-time joint semantic reasoning for autonomous driving. In *2018 IEEE Intelligent Vehicles Symposium*, pages 1013–1020. IEEE, 2018.

[226] Wen Liu, Yankui Sun, and Qingge Ji. Mdan-unet: Multi-scale and dual attention enhanced nested u-net architecture for segmentation of optical coherence tomography images. *Algorithms*, 13(3):60, 2020.

[227] Stefanos Apostolopoulos, Carlos Ciller, Sandro De Zanet, Sebastian Wolf, and Raphael Sznitman. Retinet: Automatic amd identification in oct volumetric data. *Investigative Ophthalmology & Visual Science*, 58(8):387–387, 2017.

[228] Jeroen Van der Laak, Geert Litjens, and Francesco Ciompi. Deep learning in histopathology: the path to the clinic. *Nature medicine*, 27(5):775–784, 2021.

[229] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

[230] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019.

[231] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[232] Johann Radon. 1.1 über die bestimmung von funktionen durch ihre integralwerte längs gewisser mannigfaltigkeiten. *Classic papers in modern diagnostic radiology*, 5(21):124, 2005.

[233] Caroline Richmond. Sir godfrey hounsfield. *BMJ : British Medical Journal*, 329(7467):687, 2004.

[234] Paul C Lauterbur. Image formation by induced local interactions: examples employing nuclear magnetic resonance. *nature*, 242(5394):190–191, 1973.

[235] WH Escovitz, TR Fox, and R Levi-Setti. Scanning transmission ion microscope with a field ion source. *Proceedings of the National Academy of Sciences*, 72(5):1826–1828, 1975.

[236] Jonathan H Orloff and Lynwood W Swanson. Study of a field-ionization source for microprobe applications. *Journal of vacuum science and technology*, 12(6):1209–1213, 1975.

[237] Olivier Chapelle and Alexander Zien. Semi-supervised classification by low density separation. In *International workshop on artificial intelligence and statistics*, pages 57–64. PMLR, 2005.

[238] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos

Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.

[239] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[240] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

[241] Xiaomeng Li, Lequan Yu, Hao Chen, Chi-Wing Fu, Lei Xing, and Pheng-Ann Heng. Transformation-consistent self-ensembling model for semisupervised medical image segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(2):523–534, 2020.

[242] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR, 2016.

[243] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *Advances in neural information processing systems*, 29, 2016.

[244] Tianfei Zhou, Wenguan Wang, Ender Konukoglu, and Luc Van Gool. Rethinking semantic segmentation: A prototype view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2582–2593, 2022.

[245] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[246] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on*

computer vision and pattern recognition (CVPR'06), volume 2, pages 1735–1742. IEEE, 2006.

[247] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

[248] Zhi Tian, Chunhua Shen, Xinlong Wang, and Hao Chen. Boxinst: High-performance instance segmentation with box annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5443–5452, 2021.

[249] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *Proceedings of machine learning and systems*, 2:129–146, 2020.

[250] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari S. Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *arXiv preprint arXiv:2206.14486*, 2022.

[251] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[252] Guotai Wang, Maria A Zuluaga, Wenqi Li, Rosalind Pratt, Premal A Patel, Michael Aertsen, Tom Doel, Anna L David, Jan Deprest, Sébastien Ourselin, et al. Deepigeos: a deep interactive geodesic framework for medical image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1559–1572, 2018.

[253] Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer, 2018.

[254] Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance - how to properly compare deep neural models. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics, 2019.

[255] Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. deep-significance: Easy and meaningful signifcance testing in the age of neural networks. In *ML Evaluation Standards Workshop at the Tenth International Conference on Learning Representations*, 2022.

[256] Donghuan Lu, Morgan Heisler, Sieun Lee, Gavin Ding, Marinko V Sarunic, and Mirza Faisal Beg. Retinal fluid segmentation and detection in optical coherence tomography images using fully convolutional neural network. *arXiv preprint arXiv:1710.04778*, 2017.

[257] Constantin Seibold, Simon Reiß, Jens Kleesiek, and Rainer Stiefelhagen. Reference-guided pseudo-label generation for medical semantic segmentation. *arXiv preprint arXiv:2112.00735*, 2021.

[258] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[259] Chaz Firestone and Brian J Scholl. "please tap the shape, anywhere you like" shape skeletons in human vision revealed by an exceedingly simple measure. *Psychological science*, 25(2):377–386, 2014.

[260] Shuwei Zhai, Guotai Wang, Xiangde Luo, Qiang Yue, Kang Li, and Shaoting Zhang. Pa-seg: Learning from point annotations for 3d medical image segmentation using contextual regularization and cross knowledge distillation. *arXiv preprint arXiv:2208.05669*, 2022.

[261] Larissa Heinrich, Jan Funke, Constantin Pape, Juan Nunez-Iglesias, and Stephan Saalfeld. Synaptic cleft segmentation in non-isotropic volume electron microscopy of the complete drosophila brain. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 317–325. Springer, 2018.

[262] Philipp Kainz, Martin Urschler, Samuel Schulter, Paul Wohlhart, and Vincent Lepetit. You should use regression to detect cells. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 276–283. Springer, 2015.