Graph-constrained Contrastive Regularization for Semi-weakly Volumetric Segmentation

Simon Reiß[1], Constantin Seibold[1], Alexander Freytag[2], Erik Rodner[3], and Rainer Stiefelhagen[1]

[1] Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany
[2] Carl Zeiss AG, 07745 Jena, Germany
[3] University of Applied Sciences Berlin, 12459 Berlin, Germany
{simon.reiss, constantin.seibold, rainer.stiefelhagen}@kit.edu,
alexander.freytag@zeiss.com, erik.rodner@htw-berlin.de

Abstract. Semantic volume segmentation suffers from the requirement of having voxel-wise annotated ground-truth data, which requires immense effort to obtain. In this work, we investigate how models can be trained from *sparsely annotated volumes*, *i.e.* volumes with only individual slices annotated. By formulating the scenario as a semi-weakly supervised problem where only some regions in the volume are annotated, we obtain surprising results: expensive dense volumetric annotations can be replaced by cheap, partially labeled volumes with limited impact on accuracy *if* the hypothesis space of valid models gets properly constrained during training. With our *Contrastive Constrained Regularization (Con2R)*, we demonstrate that 3D convolutional models can be trained with less than 4% of only two dimensional ground-truth labels and still reach up to 88% accuracy of fully supervised baseline models with dense volumetric annotations. To get insights into *Con2R*s success, we study how strong semi-supervised algorithms transfer to our new volumetric semi-weakly supervised setting. In this manner, we explore retinal fluid and brain tumor segmentation and give a detailed look into accuracy progression for scenarios with extremely scarce labels.

Keywords: Volumetric semantic segmentation, semi-weakly supervised learning, regularization, contrastive learning

1 Introduction

Over the last decades, healthcare and natural sciences underwent a drastic increase in efficiency in analyzing data by exploiting semantic segmentation algorithms. Not surprisingly, current products and solutions in these domains are built with neural networks, as their performance has proven superior in most use cases [1,7,16]. And still, one major challenge often limits scaling applications further: the need for precisely annotated data. This especially holds for segmentation networks to generalize well to new examples – which in these applications is not only of academic interest, but of utmost importance. However, *e.g.* in healthcare, only trained experts are able to provide correct annotations.

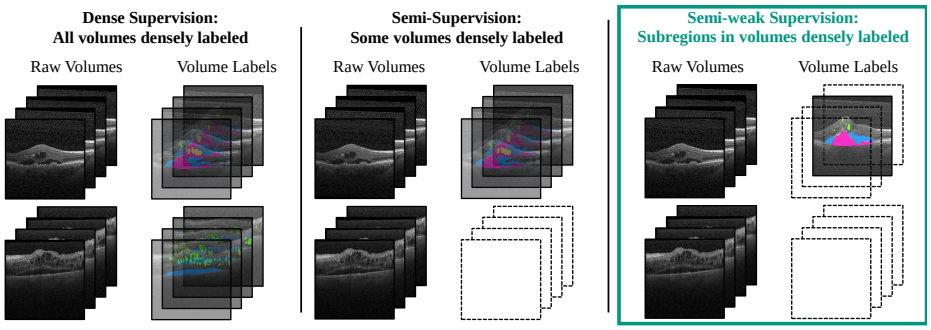


Fig. 1. In which setting can 3D segmentation become relevant in practise? *Left*: densely supervised volume segmentation requires fully annotated volumes – hardly possible in practise. *Center*: semi-supervision allows some volumes to be unlabeled – better, but still with fully annotated volumes. *Right*: we propose semi-weakly volume segmentation as the missing step for bringing 3D segmentation to practise at affordable costs

These experts are hard to get for annotation tasks, as their main job is often not less but to save lives. Constructing large labeled datasets therefore becomes expensive and difficult. Even worse, the wide-spread use of volumetric image data with all of its benefits, *e.g.* as prevalent in computed tomography [25], optical coherence tomography [7], or magnetic resonance imaging [1,54], amplifies this annotation problem by going from 2D pixel- to even more laborious 3D voxel annotations (if you do not agree with this sentence, you should spend at least one hour on annotating volume data yourself).

These observations clearly show the urgent need for training schemes for 3D segmentation networks to become more economic and more annotation-efficient. We will therefore investigate how these models can be trained with partially labeled as well as entirely unlabeled volumes. We aim at answering the question *Can we circumvent the additional effort to annotate entire volumes?* Partially labeled volumes, *e.g.* volumes that have at most few individual image-slices annotated, can be considered weak labels for the 3D volumetric segmentation task (Fig. 1). As we train with weak annotations and cheaper unlabeled volumes, we refer to it as semi-weakly supervised volumetric segmentation (as in weakly- and semi-supervised [13,50]). Our contributions can be summarized as follows:

- We establish the task of semi-weakly supervised volumetric segmentation and set up thorough training and evaluation protocols for it.
- We analyze and transfer established semi-supervised methods from 2D and 3D to the semi-weakly supervised volume segmentation setting, which gives strong baselines and insights into performance implications.
- We propose the *Con2R* objective for training volumetric models on sparsely labeled data, integrating *smoothness-* and *semantic coherence constraints*.
- By considering the mismatch in training and testing targets, we achieve performances of up to 88% as compared to densely supervised models on the RETOUCH dataset with merely two dozen labels (a fraction of 3.5% labels) and outperform all semi-supervised models on BraTS.

2 Related Work

Semantic segmentation In semantic segmentation literature steep progress on commonly known natural image benchmarks [17,20,41,71] has been made. Impressive performance is achieved using convolutional [11,27] or transformer-based architectures [56,67] and novel training strategies [4,74]. These models are often trained or pre-trained on gigantic datasets [14,26,74], to itch out every performance improvement, which for domains that are more distant from natural imagery and do not encompass this data-richness, often lies beyond reach. This might be one reason that arguably one of the most successful architectures for segmentation in domains distant from ImageNet [53] is the UNet architecture [51] and its 3D counterpart [15] for volumetric data. In our work, we are interested in domains distant from natural imagery and mainly explore medical data [1,7] where limited amounts of semantic labels are available.

Volumetric semantic segmentation We are further interested in processing volumetric data, which started in the neural network era with video processing [32,34,58] and was shortly after adapted to voxel-wise prediction tasks [35,59]. With their introduction of the 3D UNet Çiçek *et al.* [15] set of a string of works that culminate in the state-of-the-art volumetric segmentation architectures as indicated by 3D approaches [30,31,43,45,46,69] dominating leaderboards of common benchmarks [3,25]. A lot of flavors of the 3D UNet have been proposed: adding multiple pathways [33], deep supervision [63], self-supervised training objectives [46] as well as specifically considering boundaries [24,33,63] . We lay emphasis on exploring how 3D models can cope with weaker training signals than dense volume annotations. Closest to our work are [18], a 3D segmentation model trained on retinal OCT scans, which is done partly on sparse labels as in [15] and [50] which shares the very low-data regimen for retinal fluid segmentation.

Semi- and weak volume supervision Coping with fewer labels for volumetric segmentation has seen a lot of interest recently and was most commonly posed as semi-supervised task, *i.e.* labeled and unlabeled volumes are used for training. For this, a variety of approaches were tested, *e.g.* based on adversarial learning [47], integrating the 3D shape of the input data by training distinctly on multiple views and fuse model predictions on unlabeled volumes [72].Mean-Teachers [57] were also successfully applied in volumetric segmentation [61,62,66] often in combination with uncertainty modeling. Uncertainty is also integrated in [42], yet rather than previous approaches which leverage classical Monte Carlo Dropout [22] they base the uncertainty measure on predictions from multiple UNet decoding scales and directly minimize it. Aside from designing new training strategies by splitting the training data into different sets [28], contrastive learning has seen some application in semi-supervised volume segmentation recently [64,65,68]. In these works, contrastive learning is used on a voxel-level [64] and slice-level [65,68], which is coarsely related to our work through the idea of enforcing similarities. Yet, we don't set up positives and negatives, we deliberately design target voxel-similarities based on positional- and semantic proximity.

A lot of semi-supervised algorithms consider graphs [48,73], while many graphs are built between labeled and unlabeled samples [29,38], we bring this

view to the voxel level within individual volumes. Comatch [38] considers semi-supervised contrastive learning as graph-regularization for 2D and builds embedding graphs across different images, rather than different voxel-embeddings. Our method also coarsely relates to segmentation post-processing methods such as CRFs [8,10,36,70], which view individual images as graphs and pixels as vertices.

We restrict our supervision types to unlabeled- alongside sparsely annotated volumes, which is related to weakly supervised literature [5,40]. In the medical domain, scribbles which are also partial labels, were used in [60] to learn 2D segmentation via an adversarial objective and access to unpaired dense labels. Singular points as partial annotations were explored in [49], where a segmentation model is bootstrapped by iteratively pseudo-labeling unlabeled regions of histopathology images. For volumetric segmentation, labeling only extreme points on the three dimensional entity to segment built the foundation of [52], which, like our method, views input volumes as graphs.

3 Proposed Approach

In this chapter, we introduce our notation, define the task of semi-weakly supervised volumetric segmentation, and discuss relevant network related architectural choices. Then, we outline our *Contrastive Constrained Regularization (Con2R)* method that can be understood as graph constraints on the learned feature space. Via its design as contrastive loss, it encompasses a *receptive field smoothness constraint* and a *semantic coherence constraint*.

3.1 Preliminaries

Supervision modality and notation We leverage a volume dataset of size N for semantic segmentation, with the specification of:

$$\mathcal{D} = \{v_1, ..., v_N | v_i \in \mathbb{R}^{c_{dim} \times D \times H \times W}\} \ , \tag{1}$$

where c_{dim} is the number of volume input channels, its depth D, height H, and width W. In the general setting of volumetric segmentation, an input volume v_i is accompanied by a ground-truth $m_i \in \mathbb{R}^{C \times D \times H \times W}$, with C classes to segment. Our setting reduces the requirement for densely labeled ground-truths as follows:

$$\mathcal{M} = \{(m_1, a_1), ..., (m_N, a_N) | (m_i, a_i) \in (\mathbb{R}^{C \times D \times H \times W}, \{0,1\}^D)\} \ . \tag{2}$$

In addition to ground-truth annotations m_i, we use binary variables $a_i^d \in \{0,1\}$ to indicate the availability of annotation information at each slice location $1 \leq d \leq D$ in a volume v_i. An indicator a_i that only contains zeros corresponds to m_i not containing any annotation information, whereas an indicator a_i containing *e.g.* two ones indicates the locations of two annotated slices within m_i.

Generally, we care about the setting where indicators satisfy $\sum_{d=1}^{D} a_i^d \ll D$, *i.e.* v_i are very sparsely annotated. We further speak of semi-weakly supervised

volume segmentation if these sparse annotations are distributed among all volumes in the dataset: $\sum_{i,d=1}^{N,D} a_i^d \ll N \cdot D$. The goal for a learning algorithm stays consistent with traditional volumetric segmentation, given any unseen v, predict the full corresponding semantic m.

We found that this scenario is important in practice when experts are asked to add pixel-wise semantic labels to slices within volumetric data. The question this formulation poses is: Can three dimensional segmentation also be performed well when only weak two dimensional annotations are available?

Volume indexing and -processing Given an input volume v, we refer to v^x as the input-values at voxel location x. Throughout our experiments, we leverage 3D segmentation architectures that produce voxel-wise features $f \in \mathbb{R}^{f_{dim} \times D \times H \times W}$ and voxel-wise semantic predictions $p \in \mathbb{R}^{C \times D \times H \times W}$ (for instance 3D Encoder-Decoder models [15]). Here, the predictions $p = \kappa(f)$ are the result of the output-head $\kappa(\cdot)$, which is parameterized by C $1 \times 1 \times 1$ convolution kernels. Similarly, our method processes the voxel-wise features and transforms them via $\tau(f)$ into embeddings $e \in \mathbb{R}^{e_{dim} \times D \times H \times W}$. The transformation function $\tau(\cdot)$ is parameterized by a sequence of: normalization layer, $1 \times 1 \times 1$ convolution, non-linearity (*i.e.* LeakyReLU) and a final $1 \times 1 \times 1$ convolution layer. With this dual-head 3D segmentation architecture, each voxel v^x from the input volume can be described by a semantic prediction p^x as well as a high dimensional voxel-embedding e^x.

3.2 Graph Constraints as Regularization

In the absence of densely labeled data, we fall back to designing data-driven constraints on the hypothesis space to address the train-test target mismatch to be suitable. In particular, we take the commonly chosen view on the input data as a graph as also done in the excellent papers [10,36,38,70].

Our method considers a complete bi-partite weighted graph $\mathcal{G} = (\mathcal{Q}, \mathcal{N}, \mathcal{E}, \sigma)$. In this graph, we have two sets of vertices, the Query-set $\mathcal{Q}$ and Neighborhood-set $\mathcal{N}$ which both contain voxel-embeddings e^x sampled from all voxel-embeddings e in the current volume (*w.l.o.g.* we choose $|\mathcal{Q}| = |\mathcal{N}|$). Important note: we do not restrict these sets to local neighbors in the volume, but instead sample globally from all possible pairs of voxels during training. The vertices, *i.e.* embeddings $e^x \in \mathcal{Q}$, are connected to embeddings in $e^y \in \mathcal{N}$ by edges $(x, y) \in \mathcal{E}$ and weighted by $\sigma(x, y) = {e^x}^T e^y / (||e^x|| ||e^y||)$. Thus, our graph $\mathcal{G}$ describes the similarity between voxel-embeddings of a volume.

We can now exploit this graph to regularize model training by preferring solutions where the weights in the differentiable graph $\mathcal{G}$ take specific target values. The question remains: what are sensible choices for a function $\mathcal{T}(\cdot)$ to define target values for each edge? In the standard case of full supervision, $\mathcal{T}(\cdot)$ can simply set the weights in $\mathcal{E}$ to class-agreement between the vertices based on their labels. As we lack this option, we need to design these targets on the basis of practical assumptions that integrate knowledge about the relationship between unlabeled and labeled data [6].

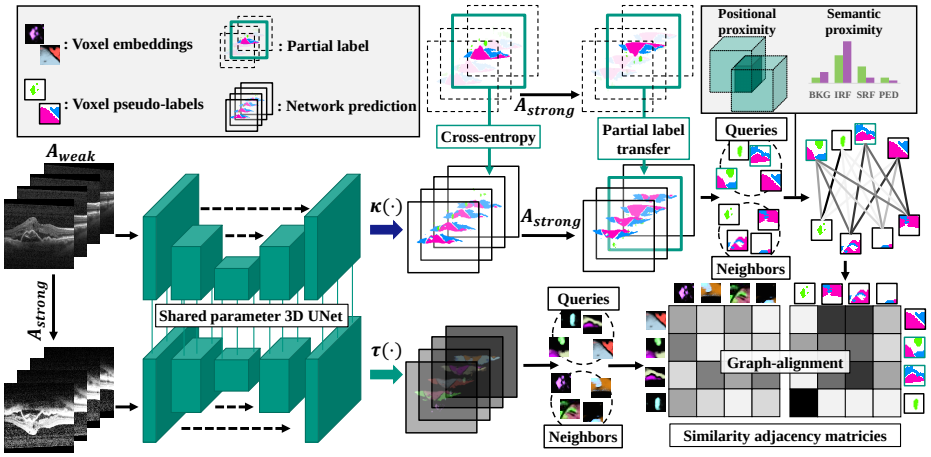


Fig. 2. Our method *Con2R* processes weakly- and strongly augmented volumes to generate voxel-wise embeddings and form a similarity graph. We align this graph to a target similarity graph that we compute via positional- and semantic proximity constraints using network predictions and partial labels if available. This alignment process enables us to learn consistent 3D predictions, merely using unlabeled or partially labeled data.

Receptive Smoothness Constraint The well-known smoothness assumption [9], which states that *samples close to each other likely share a class label*, is the basis of much recent work [21,57], which enforce consistent predictions between differently perturbed versions of the same input. Similar but differently, we further condition the smoothness assumption on the magnitude of such perturbations: *samples closer to each other are more likely to share a class label*. With the similarity graph-based design introduced above, we are now able to integrate this assumption easily. By considering translations as a form of perturbation, we can enforce the similarity between voxel-embeddings to be proportional to their relative position in the volume. Thus, we condition the smoothness assumption on the magnitude of the translation, *i.e.* on the relative position.

To integrate this assumption into our target similarity weights, let us consider two embeddings $e^x \in \mathcal{Q}$ and $e^y \in \mathcal{N}$. We propose to compute *positional proximity* of the two voxel-embeddings in the volume by using the relative intersection of sub-volumes centered at x and y (smoothed by a small ε if the intersection approaches 0):

$$\rho(x, y, \mathcal{R}(\cdot)) = \max\left(\frac{|\mathcal{R}(x) \cap \mathcal{R}(y)|}{|\mathcal{R}(x)|}, \varepsilon\right) \ , \tag{3}$$

where the receptive field function $\mathcal{R}(\cdot)$ returns for a voxel x all spatially related voxels that fall into the sub-volume centered at x. For a simplified depiction (2D case), see Fig. 3, where in (2) voxel B shares a larger receptive portion with A than with C or D, therefore B's voxel-embedding should be *closest* to A's. This positional similarity is marginalized over neighborhood embeddings $e^z \in \mathcal{N}$:

$$\mathcal{P}(x, y, \mathcal{R}(\cdot)) = \frac{\rho(x, y, \mathcal{R}(\cdot))}{\sum_{e^z \in \mathcal{N}} \rho(x, z, \mathcal{R}(\cdot))} \ . \tag{4}$$

When no semantic information on class membership is present, constraining a model to be coherent with respect to $\mathcal{P}(\cdot)$ results in models that draw conclusions concerning embedding-similarity exclusively on the basis of relative positioning. In this case, the resulting models will lead to embeddings which consider the full extent of three dimensional receptive volumes. As it will turn out, this is useful in our setting with extremely few annotated volume slices.

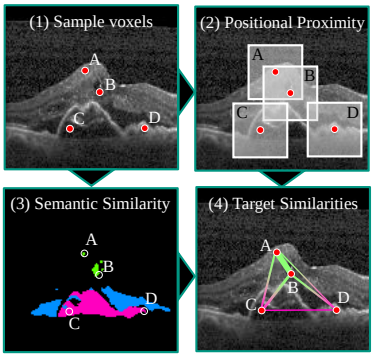


Fig. 3. Graph constraints (simplified 2D): Pairs of voxels are related by positional proximity measured in overlap of receptive field/volume and similarity in class-predictions.

Semantic Coherence Constraint Besides positional proximity, we also enforce coherence of the embeddings for voxels with similar semantics. This is important to offset the position constraint for voxel-embeddings that share semantics but lie far apart, *e.g.* voxel C and D in (3) of Fig. 3 share the class prediction, hence C's voxel-embedding should be *more similar* to D than to A or B. For two embeddings $e^x \in \mathcal{Q}$ and $e^y \in \mathcal{N}$, we take into account the semantic predictions p^x and p^y produced by the segmentation output-head $\kappa(\cdot)$, which is trained with the few given labels. To measure *semantic proximity*, different functions $\mathcal{S}(\cdot)$ have been proposed [36]. We base our measure on the symmetrized negative Kullback-Leibler divergence:

$$\text{SN-KL}(p^x, p^y) = -\frac{1}{2} \cdot \left(p^y \cdot \log\left(\frac{p^y}{p^x}\right) + p^x \cdot \log\left(\frac{p^x}{p^y}\right)\right) \quad , \tag{5}$$

which we marginalize over the predictions at the locations of the neighborhood voxels:

$$\mathcal{S}(x, y, p) = \frac{\exp(\text{SN-KL}(p^x, p^y))}{\sum_{e^z \in \mathcal{N}} \exp(\text{SN-KL}(p^x, p^z))} \quad . \tag{6}$$

With the semantic proximity $\mathcal{S}(\cdot)$ and the positional proximity $\mathcal{P}(\cdot)$, we are now able to set up the function $\mathcal{T}(\cdot)$, which produces target similarity-weights for $\mathcal{G}$.

Graph-based Contrastive Constraints To restrict the similarity weights of our embedding graph $\mathcal{G}$, we first define a function $\mathcal{T}$ to obtain the target similarities between pairs of voxel-embeddings. For a given edge (x, y) between voxel-embeddings, the model should produce a similarity $\sigma(x, y)$ that matches:

$$\mathcal{T}(x, y, \mathcal{R}(\cdot), p) = \alpha \cdot \mathcal{P}(x, y, \mathcal{R}(\cdot)) + (1 - \alpha) \cdot \mathcal{S}(x, y, p) \quad . \tag{7}$$

The weight $\alpha \in [0, 1]$ allows to trade-off the contribution of the *receptive smoothness-* and *semantic coherence constraints*. With $\mathcal{T}$, we can align the voxel-embeddings for a given volume to these similarity targets. For this alignment process, we leverage the common contrastive similarity formulation:

$$\mathcal{O}(e^x, e^y) = \frac{\exp(\sigma(x, y))}{\sum_{e^z \in \mathcal{N}} \exp(\sigma(x, z))} \quad , \tag{8}$$

which encodes the current voxel-embedding similarities in the graph $\mathcal{G}$. The final loss is formed by minimizing the cross-entropy between similarities and targets:

$$L(\mathcal{Q},\mathcal{N}) = -\sum_{e^x\in\mathcal{Q},e^y\in\mathcal{N}} \log(\mathcal{O}(e^x,e^y))\cdot\mathcal{T}(x,y,\mathcal{R}(\cdot),p) \ . \tag{9}$$

Finally, we symmetrize this loss and follow the example of [12] by only backpropagating through either $\mathcal{Q}$ or $\mathcal{N}$. Therefore, our proposed L_{Con2R} loss function resolves to:

$$L_{Con2R}(\mathcal{Q},\mathcal{N}) = \frac{1}{2}\cdot(L(\mathcal{Q},\bar{\mathcal{N}}) + L(\mathcal{N},\bar{\mathcal{Q}})) \ . \tag{10}$$

Here, $\bar{\mathcal{Q}}$ and $\bar{\mathcal{N}}$ indicate the respective voxel-embedding sets being detached from the computation graph, *i.e.* we treat them as constants in backpropagation.

3.3 Graph-constrained Semi-weak Learning

To put L_{Con2R} to work, we propose the following training scheme which is also visualized in Figure 2. First, we take an input volume v_i and augment it weak and strong, yielding $\mathcal{A}_{weak}(v_i)$ and $\mathcal{A}_{strong}(v_i)$. We train the output-head $\kappa(\cdot)$ with only the weakly augmented volumes by minimizing standard categorical cross-entropy $L_{Entropy}$ using the few partially labeled annotations. We use the predictions p from the weakly augmented volumes as input to generate our target similarities $\mathcal{T}(\cdot)$. When sparse annotations for v_i are present, we adapt p_i to p_i^*:

$$p_i^* = p_i\cdot(1-a_i) + m_i\cdot a_i \ , \tag{11}$$

substituting ground-truth annotations in regions of the volume, where we are supplied with them. The embeddings we use for setting up the similarity graph $\mathcal{G}$ are taken from the strongly augmented input, *i.e.* embeddings produced by forwarding $\mathcal{A}_{strong}(v_i)$ through the network as well as the embedding output-head $\tau(\cdot)$. Put together, we optimize our semi-weakly supervised 3D segmentation model by minimizing $L_{total} = L_{Entropy} + L_{Con2R}$.

4 Evaluation

4.1 Protocol

Datasets We evaluate our approach on two well-known volumetric datasets. The RETOUCH OCT dataset [7] for retinal fluid segmentation contains classes: Intraretinal fluid (IRF), Subretinal fluid (SRF), and Pigment Epithelium Detachment (PED). While different vendors of OCT devices are covered, we focus on Spectralis, for which the volumes have a depth of 49 B-scans. Further, we evaluate our approach on brain tumor sub-region segmentation in magnetic resonance images. We use the data as supplied in [1] which contains multiple BraTS challenges [2,3,44]. We segment tumor sub-regions edema (EDM), enhancing tumor (EN), non-enhancing tumor (NEN) within volumes of depth 155.

Experimental setup We intend to analyze extremely scarce annotation scenarios. Thus, we need to carefully design a suitable evaluation protocol. Therefore, we split the labeled data five times into train and test splits, where the train portion is further divided into train and validation volumes (train/val/test: RETOUCH: 14/10/10, BraTs: 242/121/121). This five-fold cross-validation enables us to report mean and standard deviation for the performance of all models which is crucially important when working with few labels [50]. In each train-val fold, we randomly select B-scans to be annotated, thereby marginalizing effects of individual annotated slices. We make sure that the set of annotated scans covers all classes. As such, in each fold a sequence of scenarios where only $3, 6, 12$, or 24 label-masks are available gives detailed insight into effects of adding annotations. With *e.g.* $\mathcal{M}(12)$ we refer to a scenario that from all training volumes only has access to 12 annotated slices, distributed among all volumes ($\sum_{i,d=1}^{N,D} a_i^d = 12$). In a fold, higher supervision scenarios extend labels of lower supervision scenarios.
Evaluation metric For evaluating the performance of our methods, we use the mean Intersection over Union (mIoU), which is defined as the average class-wise Intersection of segmentation and ground-truth over their Union. We report mIoU averaged over five cross-validation folds and the standard deviation.

4.2 Implementation details

Data augmentation. For RETOUCH, we resize the input volumes to $49 \times 160 \times 160$ and crop out 16 slices for training, while for BraTS, we resize to $155 \times 110 \times 110$ and crop out $32 \times 110 \times 110$-sized volumes. As described in Section 3.3, we require weak and strong augmentations for computing pseudo-labels and voxel-embeddings. Weak augmentations are in our setting flipping the input volume in the longitudinal- and the vertical direction with a probability of 50%. To compute embeddings, the input is flipped in all three directions with a probability of 50% and always altered via photometric perturbations with sampled magnitudes. Here, we find adjusting the brightness and sharpness to be most effective. Furthermore, we extend CutOut [19] for volumetric inputs, where we always set a randomly placed cube of size $16 \times 16 \times 16$ in the volume to zero.

Network configuration. For all experiments and baselines, we leverage 3D UNets [15] with $64, 128, 256, 256$ channels in each encoder and decoder layer. We train the networks using a batchsize of two, where we oversample the partially labeled volumes and ensure that in each iteration, one of the volumes is partially labeled. We apply Xavier initialization [23], use a learning rate of 0.01, and use SGD with momentum of 0.9 and a weight decay of 0.00001. As lower bound on performance, we train models that merely employ cross-entropy on partially labeled volumes and do not consider unlabeled volumes or unlabeled regions. These naive 3D UNet baselines serve as initialization to the semi-supervised models. Training is conducted in 100 epochs and validated every 10 epochs, where the best epoch is then evaluated once on the testing set. All experiments were carried out on 11GB NVIDIA RTX 2080 Ti GPUs. The code is made publicly available at `https://github.com/Simael/Con2R`.

4.3 Baselines and Setup for Semi-weak Volumetric Segmentation

We implement common methods from semi-supervised literature and tune them for semi-weakly volumetric segmentation. Thus, we select the most common [37] and successful 2D methods [55] and methods that have seen success on volumetric medical use-cases [39,66] and also a naive **3D UNet** as a lower bound. With this we explore how they compare and transfer to partially labeled scenarios.
Pseudo-label (PL) [37,74] We implement a pseudo-labeling baseline which is a common, classical semi-supervised approach. As we have access to partially labeled volumes, in pseudo-labeling we use the network predictions and augment them via Eq. (11). This simple transfer of the method did not perform well in our setting; we tune it by using the self-training normalization scheme of [74].
Mean-Teacher (MT) [57] A commonly used semi-supervised framework is the Mean-Teacher, which we transfer to volumetric inputs and dense predictions. The adaptation of this approach to segmentation has been modeled multiple times [39,50]. We train by aligning the student predictions to the teacher predictions as obtained by forward passing differently augmented volumes. Then, we reverse geometric augmentations on the teacher predictions to maintain pixel-alignment between student- and teacher outputs for the consistency loss. We find an exponential-moving average decay factor 0.5 to perform well.
FixMatch [55] FixMatch is a successful method originally designed for 2D classification. It mixes pseudo-labeling and consistency regularization by using weak and strong augmentations (we use augmentations from Sec. 4.2). As we adapt this approach to segmentation, we consider the alignment of predictions from the strongly augmented branch to the weakly augmented branch, which we do similarly as in the Mean-Teacher. A confidence threshold of 0.5 was suitable.
Uncertainty-aware Mean-Teacher (UA MT) [66] This Mean-Teacher flavor adds uncertainty estimation using Monte-Carlo dropout [22]. By estimating and thresholding voxel-wise uncertainty, the consistency loss is applied selectively in unlabeled regions. We find a threshold of 0.5 and 8 forward passes to work well.
Contrastive Constrained Regularization (Con2R) We train our *Con2R* method on RETOUCH by sampling $|\mathcal{Q}| + |\mathcal{N}| = 3456$ voxels from the volume-graph (BraTS: 6750) each iteration and optimize the alignment to our computed target graph. The composition of positional- and semantic constraints for the target graph is moderated 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, we set $e_{dim} = 64$ and $\varepsilon = 10^{-7}$.

4.4 Quantitative results

RETOUCH When looking at Table 1, the first observation is that the lower accuracy bound for scenarios with $3, 6, 12, 24$ annotations is set by the 3D UNet. In the lowest supervision scenario $\mathcal{M}(3)$, results are as expected very poor and most semi-supervised models can not meaningfully exceed the plain 3D UNet. With additional supervision, semi-supervised methods start to show improvements due to modeling concistency in unlabeled data. Interestingly, FixMatch gives smallest gains in $\mathcal{M}(6)$ while in later scenarios, it is comparable to competing

Table 1. 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

Method	$\mathcal{M}(3)$	$\mathcal{M}(6)$	$\mathcal{M}(12)$	$\mathcal{M}(24)$	$\mathcal{M}(\textbf{full})$
3D UNet [15]	12.0 ± 5.6	18.1 ± 11.5	31.1 ± 12.4	43.8 ± 2.5	54.9 ± 0.9
PL [37,74]	13.0 ± 6.3	20.6 ± 13.4	30.9 ± 11.5	45.7 ± 2.2	55.4 ± 1.5
MT [39,57]	12.0 ± 6.6	20.2 ± 12.4	34.4 ± 11.4	45.3 ± 3.1	53.4 ± 1.9
FixMatch [55]	10.4 ± 5.7	18.7 ± 10.6	34.7 ± 6.8	46.2 ± 3.8	54.4 ± 3.3
UA MT [66]	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

approaches. *Con2R* is able to outperform all methods with clear margins, especially Uncertainty-aware Mean-Teacher and FixMatch, which are the strongest competitors in $\mathcal{M}(12)$ and $\mathcal{M}(24)$. With full annotations, the baselines and *Con2R* achieve, within a small margin, comparable results, as no additional unlabeled data is leveraged. UA MT further improves $\mathcal{M}(\textbf{full})$ slightly, which might be due to the integrated dropout layers. The results show that semi-supervised methods generally work for the proposed semi-weakly supervised learning with partially labeled volumes. Yet, by explicitly modeling the properties of partially labeled data with our constraints in *Con2R*, we see consistent gains of +1.8%, +1.9%, +2.1%, +2.0%, as compared to the best competing methods.

We can get more nuanced insights by comparing class-wise segmentation performance in Table 2. It is evident that even with few annotations, *Con2R* is able to segment Subretinal Fluid (SRF) and Pigment Epithelium Detachments (PED) better than alternative approaches which holds true for all scenarios. In $\mathcal{M}(24)$, our method segments Subretinal Fluid with an IoU of 79.1% which is close to the best fully supervised result of 84.4% reached by Uncertainty-aware Mean-Teacher with access to a total of 686 annotated OCT B-scans.

Table 2. RETOUCH class-wise results in mIoU for semi-weakly supervised learning, number of annotated B-scans successively increased from 3 to 24

Method	$\mathcal{M}(3)$			$\mathcal{M}(6)$			$\mathcal{M}(12)$			$\mathcal{M}(24)$		
	IRF	SRF	PED	IRF	SRF	PED	IRF	SRF	PED	IRF	SRF	PED
3D UNet [15]	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 [37,74]	**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 [39,57]	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 [55]	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 [66]	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 3. BraTS class-wise results in mIoU for semi-weakly supervised learning, number of annotated B-scans is set at 24 and **full** access is shown as upper limit

Method	$\mathcal{M}(24)$				$\mathcal{M}(\textbf{full})$
	EDM	EN	NEN	AVG	AVG
3D UNet [15]	48.7	19.6	48.1	38.8 ± 3.4	51.7 ± 7.0
PL [37,74]	49.1	21.3	50.5	40.3 ± 2.5	52.2 ± 8.4
MT [39,57]	49.1	21.7	45.0	38.6 ± 4.5	53.7 ± 5.7
FixMatch [55]	50.1	**24.2**	53.1	42.4 ± 4.9	51.0 ± 6.5
UA MT [66]	49.2	22.6	51.3	41.1 ± 3.5	52.6 ± 6.0
Con2R (Ours)	**51.8**	23.9	**53.9**	$\mathbf{43.2 \pm 3.5}$	$\mathbf{54.6 \pm 7.7}$

BraTS In Table 3, we see the same methods evaluated for brain tumor segmentation. Due to generally long training times for 3D segmentation models, we report only one semi-weakly supervised setting, namely $\mathcal{M}(24)$. Here, we see that especially edema (EDM) and non-enhancing tumor (NEN) sub-regions benefit from our modeling, which leads to superior results. It is noteworthy that in this scenario, $37,486$ slices are not annotated while only 24 have associated labels. Hence, even in this highly unbalanced setting between unlabeled and labeled scans, our method is well suited and exceeds all semi-supervised baselines.

4.5 Ablation and hyperparameter sensitivity studies

We carry out all experiments in this section on the RETOUCH dataset with the 24 annotation scenario. First, we study the effect of the weight α, which interpolates between the *positional* and *semantic constraints* (Table 6). We see that semantic constraints in isolation ($\alpha = 0.0$) produce solid results, only positional constraints surprisingly too ($\alpha = 1.0$), but best results are found with $\alpha = 0.2$.
Next, we report in Table 4 how chosing the receptive field function $\mathcal{R}$ impacts accuracy. The best results are achieved with $16 \times 16 \times 16$, which is the maximum depth of the input volume crops (therefore, we adjust $\mathcal{R}$ to $32 \times 32 \times 32$ for the BraTS task). Larger receptive volume sizes degrade the performance, and we expect that the shape and size of objects to segment in a given dataset also plays an important role regarding this choice.
Finally, the number of sampled edges from the volume-graph to tune is varied in Table 5. We see that increasing this number also steadily increases the benefit of *Con2R*. We set this hyperparameter to $1,728$ for RETOUCH and to $3,375$ for BraTS, which relates to the maximum GPU capacity available to us.

4.6 Qualitative results

RETOUCH An example of the qualitative segmentation improvement between different methods while adding annotations are shown in Figure 4. With 3 annotated slices in training, none of the methods achieve satisfying results, merely

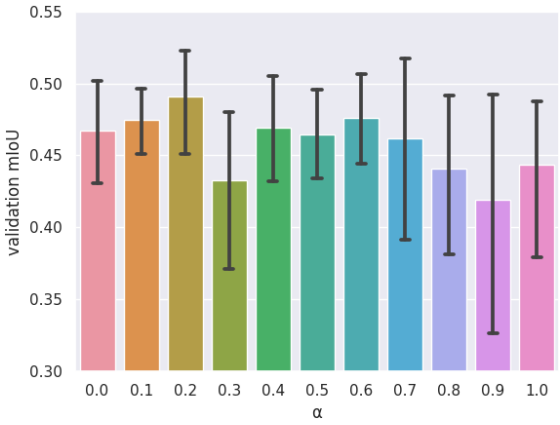
Table 4. Effect of receptive volume size $\mathcal{R}$ on the mean IoU

$\mathcal{R}$	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 number of vertices in graph $\mathcal{G}$ on the mean IoU

$\lvert\mathcal{Q}\rvert, \lvert\mathcal{N}\rvert$	validation mIoU
216	$46.9 \pm 4.5\%$
512	$46.9 \pm 3.4\%$
1000	$47.8 \pm 5.0\%$
1728	$49.1 \pm 4.7\%$

Table 6. Validation performance of *Con2R* when tuning α, IoU reported along five validation splits with mean and standard deviation displayed

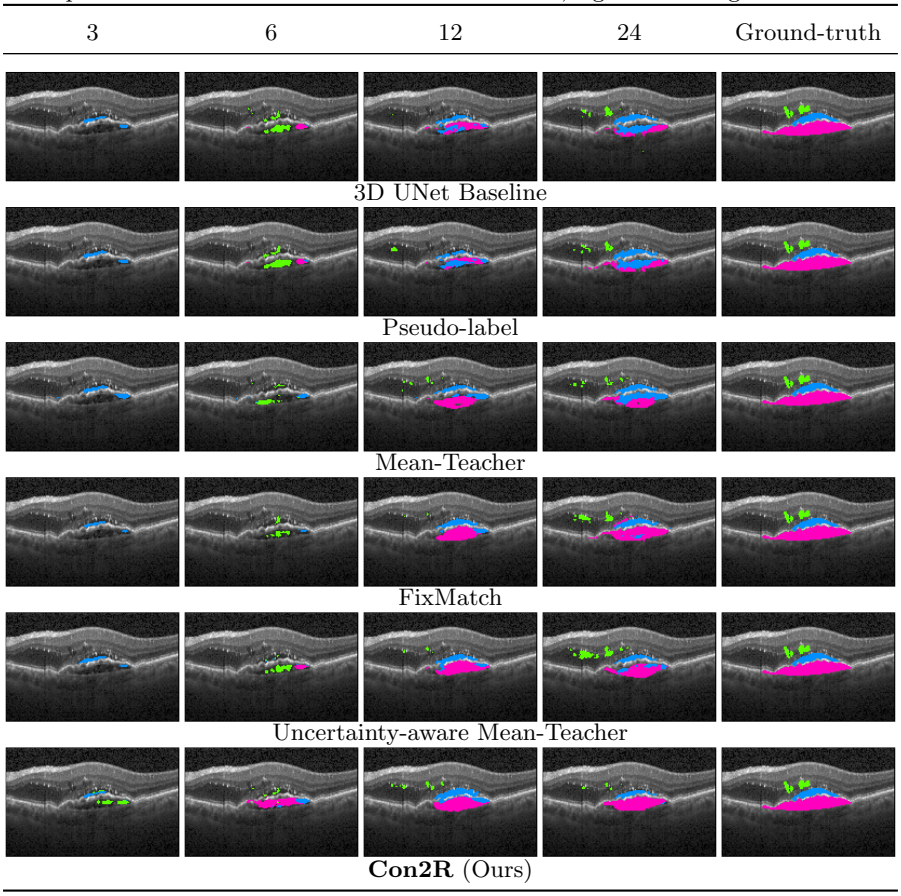


a small fluid portion of **Subretinal Fluid** (**SRF**) is in some cases segmented coarsely. Adding three annotations more, most approaches over-confidently identify **Pigment Epithelium Detachments** (**PED**) as **Intraretinal Fluid** (**IRF**), merely our method starts to segment this area correctly. FixMatch and Uncertainty-aware Mean-Teacher are the only other methods that correctly pick up the spatial relations between **PED** and **SRF** with 12 annotated slices. For this supervision scenario, our method is able to already pick up the correct location of **IRF** pockets in the retina. Using 24 annotations, our method further is the only one to correctly delineate the **SRF** and **PED** and making consistent spatial segmentations without class confusion that we see in the remaining methods. We attribute this to our smoothness priors regarding both semantics and locality.

5 Discussion

We believe that our proposed concept of semi-weakly supervised volumetric segmentation as formulated in Section 3.1 is worth exploring further since it gives detailed insights into how labeling can be optimized. Especially in application fields where labeling budget is tight or where time of expert annotators is limited, flexible learning algorithms such as *Con2R* can become enabling technologies to build useful solutions. To reduce the expert label costs further, we see potential in using annotations that only consider sub-regions within a slice in a volume. This variation from our setting would further put emphasis on intuitive expert-selected region annotation, which is one additional step towards an expert-centric process. By design, our *Con2R* method is applicable to such scenarios and we are curious to see them being analyzed in future investigations.

Fig. 4. 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. Method names are below rows, right column: ground-truth.



6 Conclusion

We introduced and explored semi-weakly volumetric segmentation to reduce the need for dense expert-labels on volumetric data. Motivated by designing flexible learning algorithms which can use partial labels, we transferred a variety of semi-supervised algorithms. It became evident that these methods indeed add performance but leave behind uncollected rewards. Our method *Con2R* recovers those by explicitly modelling the semi-weak scenario. We carefully constructed positional smoothness- and semantic coherence constraints in embedding space, and we were able to consistently raise segmentation accuracy on two medical datasets. We expect that flexible algorithms like *Con2R* which exploit unlabeled and partially labeled volume data can enable applications where annotations at scale are otherwise too costly or even impossible to obtain.

References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., van Ginneken, B., et al.: The medical segmentation decathlon. arXiv preprint arXiv:2106.05735 (2021) 1, 2, 3, 8
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017) 8
3. Bakas, S., Reyes, M., Jakab, A., Bauer, S., Rempfler, M., Crimi, A., Shinohara, R.T., Berger, C., Ha, S.M., Rozycki, M., 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) 3, 8
4. Bao, H., Dong, L., Wei, F.: Beit: Bert pre-training of image transformers. arXiv preprint arXiv:2106.08254 (2021) 3
5. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: European conference on computer vision. pp. 549–565. Springer (2016) 4
6. Ben-David, S., Lu, T., Pál, D.: Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In: COLT. pp. 33–44 (2008) 5
7. Bogunović, H., Venhuizen, F., Klimscha, S., Apostolopoulos, S., Bab-Hadiashar, A., Bagci, U., Beg, M.F., Bekalo, L., Chen, Q., Ciller, C., et al.: Retouch: the retinal oct fluid detection and segmentation benchmark and challenge. IEEE transactions on medical imaging **38**(8), 1858–1874 (2019) 1, 2, 3, 8
8. Chandra, S., Kokkinos, I.: Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In: European conference on computer vision. pp. 402–418. Springer (2016) 4
9. Chapelle, O., Zien, A.: Semi-supervised classification by low density separation. In: International workshop on artificial intelligence and statistics. pp. 57–64. PMLR (2005) 6
10. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014) 4, 5
11. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 801–818 (2018) 3
12. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021) 8
13. Choe, J., Oh, S.J., Lee, S., Chun, S., Akata, Z., Shim, H.: Evaluating weakly supervised object localization methods right. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3133–3142 (2020) 2
14. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017) 3
15. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: learning dense volumetric segmentation from sparse annotation. In: International conference on medical image computing and computer-assisted intervention. pp. 424–432. Springer (2016) 3, 5, 9, 11, 12

16. Cityscapes-Team: Semantic understanding of urban street scenes: Pixel-level semantic labeling task. https://www.cityscapes-dataset.com/benchmarks/, accessed: 2022-03-03 1
17. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016) 3
18. De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., et al.: Clinically applicable deep learning for diagnosis and referral in retinal disease. Nature medicine **24**(9), 1342–1350 (2018) 3
19. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017) 9
20. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**(2), 303–338 (2010) 3
21. French, G., Laine, S., Aila, T., Mackiewicz, M., Finlayson, G.: Semi-supervised semantic segmentation needs strong, varied perturbations. arXiv preprint arXiv:1906.01916 (2019) 6
22. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016) 3, 10
23. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics. pp. 249–256. JMLR Workshop and Conference Proceedings (2010) 9
24. Hatamizadeh, A., Terzopoulos, D., Myronenko, A.: Edge-gated cnns for volumetric semantic segmentation of medical images. arXiv preprint arXiv:2002.04207 (2020) 3
25. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., et al.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) 2, 3
26. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 3
27. Huang, S., Lu, Z., Cheng, R., He, C.: Fapn: Feature-aligned pyramid network for dense image prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 864–873 (2021) 3
28. Huo, X., Xie, L., He, J., Yang, Z., Zhou, W., Li, H., Tian, Q.: Atso: Asynchronous teacher-student optimization for semi-supervised image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1235–1244 (2021) 3
29. Iscen, A., Tolias, G., Avrithis, Y., Chum, O.: Label propagation for deep semi-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5070–5079 (2019) 3
30. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: International MICCAI Brainlesion Workshop. pp. 234–244. Springer (2018) 3
31. Isensee, F., Maier-Hein, K.H.: An attempt at beating the 3d u-net. arXiv preprint arXiv:1908.02182 (2019) 3

32. Ji, S., Xu, W., Yang, M., Yu, K.: 3d convolutional neural networks for human action recognition. IEEE transactions on pattern analysis and machine intelligence **35**(1), 221–231 (2012) 3
33. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis **36**, 61–78 (2017) 3
34. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014) 3
35. Kleesiek, J., Urban, G., Hubert, A., Schwarz, D., Maier-Hein, K., Bendszus, M., Biller, A.: Deep mri brain extraction: A 3d convolutional neural network for skull stripping. NeuroImage **129**, 460–469 (2016) 3
36. Krähenbühl, P., Koltun, V.: Parameter learning and convergent inference for dense random fields. In: International Conference on Machine Learning. pp. 513–521. PMLR (2013) 4, 5, 7
37. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013) 10, 11, 12
38. Li, J., Xiong, C., Hoi, S.C.: Comatch: Semi-supervised learning with contrastive graph regularization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9475–9484 (2021) 3, 4, 5
39. Li, X., Yu, L., Chen, H., Fu, C.W., Xing, L., Heng, P.A.: Transformation-consistent self-ensembling model for semisupervised medical image segmentation. IEEE Transactions on Neural Networks and Learning Systems **32**(2), 523–534 (2020) 10, 11, 12
40. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3159–3167 (2016) 4
41. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 3
42. Luo, X., Liao, W., Chen, J., Song, T., Chen, Y., Zhang, S., Chen, N., Wang, G., Zhang, S.: 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. pp. 318–329. Springer (2021) 3
43. McKinley, R., Meier, R., Wiest, R.: Ensembles of densely-connected cnns with label-uncertainty for brain tumor segmentation. In: International MICCAI Brainlesion Workshop. pp. 456–465. Springer (2018) 3
44. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE transactions on medical imaging **34**(10), 1993–2024 (2014) 8
45. Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y.: Segmentation of kidney tumor by multi-resolution vb-nets (2019) 3
46. Myronenko, A.: 3d mri brain tumor segmentation using autoencoder regularization. In: International MICCAI Brainlesion Workshop. pp. 311–320. Springer (2018) 3

47. Nie, D., Gao, Y., Wang, L., Shen, D.: Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 370–378. Springer (2018) 3
48. Ouali, Y., Hudelot, C., Tami, M.: An overview of deep semi-supervised learning. arXiv preprint arXiv:2006.05278 (2020) 3
49. Qu, H., Wu, P., Huang, Q., Yi, J., Yan, Z., Li, K., Riedlinger, G.M., De, S., Zhang, S., Metaxas, D.N.: Weakly supervised deep nuclei segmentation using partial points annotation in histopathology images. IEEE transactions on medical imaging **39**(11), 3655–3666 (2020) 4
50. Reiß, S., Seibold, C., Freytag, A., Rodner, E., Stiefelhagen, R.: Every annotation counts: Multi-label deep supervision for medical image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9532–9542 (2021) 2, 3, 9, 10
51. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 3
52. Roth, H., Zhang, L., Yang, D., Milletari, F., Xu, Z., Wang, X., Xu, D.: 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, pp. 42–50. Springer (2019) 4
53. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015) 3
54. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., Van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019) 2
55. Sohn, K., Berthelot, D., Li, C.L., Zhang, Z., Carlini, N., Cubuk, E.D., Kurakin, A., Zhang, H., Raffel, C.: Fixmatch: Simplifying semi-supervised learning with consistency and confidence. arXiv preprint arXiv:2001.07685 (2020) 10, 11, 12
56. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020) 3
57. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. arXiv preprint arXiv:1703.01780 (2017) 3, 6, 10, 11, 12
58. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015) 3
59. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Deep end2end voxel2voxel prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 17–24 (2016) 3
60. Valvano, G., Leo, A., Tsaftaris, S.A.: Learning to segment from scribbles using multi-scale adversarial attention gates. IEEE Transactions on Medical Imaging **40**(8), 1990–2001 (2021) 4
61. Xia, Y., Liu, F., Yang, D., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H.: 3d semi-supervised learning with uncertainty-aware multi-view co-training. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3646–3655 (2020) 3

62. Xia, Y., Yang, D., Yu, Z., Liu, F., Cai, J., Yu, L., Zhu, Z., Xu, D., Yuille, A., Roth, H.: Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation. Medical Image Analysis **65**, 101766 (2020) 3
63. Yang, X., Yu, L., Li, S., Wang, X., Wang, N., Qin, J., Ni, D., Heng, P.A.: Towards automatic semantic segmentation in volumetric ultrasound. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 711–719. Springer (2017) 3
64. You, C., Zhao, R., Staib, L., Duncan, J.S.: Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. arXiv preprint arXiv:2105.07059 (2021) 3
65. You, C., Zhou, Y., Zhao, R., Staib, L., Duncan, J.S.: Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation. arXiv preprint arXiv:2108.06227 (2021) 3
66. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019) 3, 10, 11, 12
67. Yuan, Y., Chen, X., Chen, X., Wang, J.: Segmentation transformer: Object-contextual representations for semantic segmentation. In: European Conference on Computer Vision (ECCV). vol. 1 (2021) 3
68. Zeng, D., Wu, Y., Hu, X., Xu, X., Yuan, H., Huang, M., Zhuang, J., Hu, J., Shi, Y.: Positional contrastive learning for volumetric medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 221–230. Springer (2021) 3
69. Zhang, Y., Wang, Y., Hou, F., Yang, J., Xiong, G., Tian, J., Zhong, C.: Cascaded volumetric convolutional network for kidney tumor segmentation from ct volumes. arXiv preprint arXiv:1910.02235 (2019) 3
70. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.: Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 1529–1537 (2015) 4, 5
71. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017) 3
72. Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E., Yuille, A.: Semi-supervised 3d abdominal multi-organ segmentation via deep multi-planar co-training. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 121–140. IEEE (2019) 3
73. Zhu, X., Ghahramani, Z., Lafferty, J.D.: Semi-supervised learning using gaussian fields and harmonic functions. In: Proceedings of the 20th International conference on Machine learning (ICML-03). pp. 912–919 (2003) 3
74. Zoph, B., Ghiasi, G., Lin, T.Y., Cui, Y., Liu, H., Cubuk, E.D., Le, Q.V.: Rethinking pre-training and self-training. arXiv preprint arXiv:2006.06882 (2020) 3, 10, 11, 12