Accelerated Semi-Supervised Medical Image Segmentation

İlkay Yıldız
BioSensics LLC
ilkay.yildiz@biosensics.com

February 1, 2022

Abstract

We address two challenges in 3D medical image segmentation using deep learning: limitations in image data labeled with high-quality ground-truth segments, and long training times. To this end, we take a semi-supervised learning approach, and propose a novel *comparative* unsupervised pre-training strategy prior to supervised training on ground-truth segments. This pre-training strategy imposes a global structure tailored for 3D medical images via a comparative approach: 2D slices corresponding to the same anatomical regions across different images are *more similar* to each other than other slices corresponding to different anatomical regions. Compared to existing contrastive pre-training methods, our approach reveals more fine-grained information, and reduces the computational bottleneck within the training objective. We evaluate our semi-supervised learning approach over two benchmark datasets provided by the Medical Image Segmentation Decathlon, employing the state-of-the-art UNET transformer architecture for segmentation. Our comparative strategy results in *up to* 17 *times faster* pre-training than contrastive learning. Moreover, semi-supervised segmentation initialized with comparative pre-training outperforms the prediction performances of *both* supervised learning baseline, as well as contrastive semi-supervised learning, by *up to* 11% *Dice score and* 8% *F1 score* in label-deficient settings.

Keywords 3D medical image segmentation · Semi-supervised learning · Learning from comparisons · Bradley-Terry · UNET transformer

1 Introduction

Medical imaging plays a critical role in modern medicine by revealing anatomical information with high resolution [23]. One popular medical image analysis task is *segmentation*, which is a process to detect relevant regions or structures within a medical image based on pixel intensity, texture, and anatomical or pathological knowledge [76]. To help clinicians analyze medical imaging data, it is essential to segment relevant image regions and extract their discriminative features. Unfortunately, manual execution of medical image segmentation is a highly tedious and time-consuming task for clinical experts [63, 55, 59]. As the volume and richness of medical image data is rapidly increasing due to improving healthcare facilities and acquisition technologies, manual segmentation is becoming infeasible. This motivates the need for automated segmentation tools that provide critical information by improving diagnosis accuracy and diagnostic decision explainability in computer-aided healthcare applications [63].

In the last decade, the rapid development of automated segmentation methods is strongly connected with the rising success of deep neural networks (DNNs) [42]. Existing DNN architectures for medical image segmentation are trained via two approaches: *supervised* and *semi-supervised*; we refer the reader to Appendix A for a detailed review. Supervised DNN methods dominated the literature in medical image segmentation with great prediction performance across many tasks and image modalities [49, 54, 61, 69]. Particularly, encoder-decoder DNN architectures have become state-of-the-art for this task [49], with arguably the most widely employed architecture being the UNET [60]. DNN methods further improved by *transformer* architectures [73, 39] incorporating a so-called *self-attention* module, which applies selective focus on relevant image regions to improve predictions. Transformers have been very recently unified with the prominent UNET architecture for supervised medical image segmentation [34, 7, 11]. Particularly, the UNET Transformer (UNETR) has become the state-of-the-art [34] by being particularly tailored for 3D images.

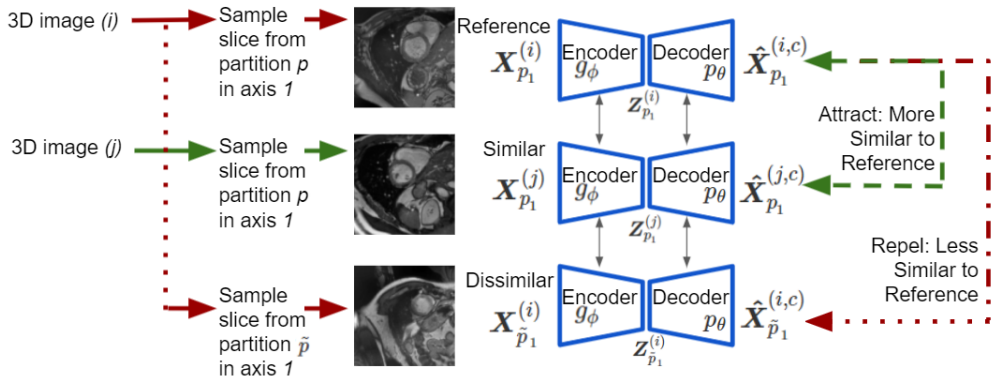


Figure 1: Our Novel Comparative Unsupervised Pre-training Strategy. For each image pair $i, j \in \{1, \ldots, M\}$, $i \neq j$, and partition pair $p, \tilde{p} \in \{1, \ldots, P\}$, $p \neq \tilde{p}$, $\boldsymbol{X}_{p_1}^{(i)}$, $\boldsymbol{X}_{p_1}^{(j)}$, and $\boldsymbol{X}_{\tilde{p}_1}^{(i)}$ form a comparative triplet, where $\boldsymbol{X}_{p_1}^{(i)}$ from partition p is more similar to $\boldsymbol{X}_{p_1}^{(j)}$ than $\boldsymbol{X}_{\tilde{p}_1}^{(i)}$ from a different partition $\tilde{p}$. For each such slice triplet, we impose the comparative global structure on the 3D segmentation predictions $\hat{\boldsymbol{X}}_{p_1}^{(i,c)}$, $\hat{\boldsymbol{X}}_{p_1}^{(j,c)}$, and $\hat{\boldsymbol{X}}_{\tilde{p}_1}^{(i,c)}$ made by the encoder-decoder DNN architecture for each segment class $c \in \{1, \ldots, C\}$. Slice images are borrowed from Chaitanya et al. [8] for illustration.

Nevertheless, supervised methods require large training datasets labeled with ground-truth segments, which are undesirable to obtain [59].

Alternatively, semi-supervised DNN methods are recent approaches in medical image segmentation that can successfully learn from small or partially-labeled datasets [32, 58, 82, 83, 1, 84]. Particularly, initializing a DNN architecture via *contrastive* pre-training prior to supervised fine-tuning on ground-truth segments incorporates global and local structures that are unique to 3D medical images [8, 79]; this approach led to significantly better segmentation predictions than prior supervised and semi-supervised methods. Despite the success in prediction performance, pre-training can significantly increase the total training time, even when implemented with highly powerful computational resources [13, 8, 79].

Motivated by these observations, we aim to improve semi-supervised DNN methods for 3D medical image segmentation in terms of learning speed. To this end, we propose a novel *comparative* unsupervised pre-training strategy for DNN architecture initialization. This pre-training strategy imposes a global structure tailored for 3D medical images via a comparative approach: 2D slices corresponding to the same anatomical regions across different images are *more similar* to each other than other slices corresponding to different anatomical regions. First and foremost, the comparative pre-training strategy employs fine-grained global structure information via a triplet of slices, i.e., a reference slice, a slice from the same anatomical region that is similar to the reference, and a slice from a different anatomical region that is dissimilar to the reference. In contrast, contrastive pre-training learns only from similar slice pairs across different images, which are simultaneously dissimilar to all other slices from different anatomical regions. Crucially, contrasting each similar slice to all other dissimilar slices introduces a computational bottleneck in the training objective, leading to much longer times for gradient-based training (c.f. Section 3).

Overall, we make the following contributions:

- We propose a novel *comparative* unsupervised training strategy to learn global structure from 3D medical images. Compared to contrastive learning, our approach reveals more fine-grained information, and reduces the computational bottleneck in the training objective.
- We employ our comparative training approach to pre-train DNN architectures for semi-supervised medical image segmentation, using the state-of-the-art UNETR architecture.
- We evaluate our semi-supervised learning approach over two benchmark datasets provided by the Medical Image Segmentation Decathlon [65]. Our comparative strategy results in *up to* 17 *times faster pre-training* than contrastive learning. Moreover, semi-supervised segmentation initialized with comparative pre-training outperforms the prediction performances of *both* supervised learning baseline, as well as contrastive semi-supervised learning, by up to 11% *Dice score and* 8% *F1 score* in label-deficient settings.

2 Problem Formulation

We consider a dataset of M 3D medical images. Each image $i \in \{1, \ldots, M\}$ is denoted as $\boldsymbol{X}^{(i)} \in \mathbb{R}^{N_1 \times N_2 \times N_3}$. Given, e.g., an MRI image $\boldsymbol{X}^{(i)}$, N_1, N_2, and N_3 correspond to the number of 2D slices along axial, coronal, and sagittal axes,

respectively. The medical image segmentation task aims to identify C segments in each image $\boldsymbol{X}^{(i)}$. For example, in brain tumor segmentation based on MRI [52], four segmentation classes that pertain to edema, non-enhancing solid core, necrotic/cystic core, and enhancing core are identified in each image. Overall, segmentation is essentially classification for each voxel, determining which of the C segments the voxel belongs to.

We employ an encoder-decoder DNN architecture for 3D medical image segmentation, inspired by their well-established performance across many tasks and image modalities [60, 20]. The encoder neural network $g_{\boldsymbol{\phi}}$ receives an image $\boldsymbol{X}^{(i)}$ and extracts latent features $\boldsymbol{Z}^{(i)} \in \mathbb{R}^{D_1 \times D_2 \times D_3}$ by gradual dimension reduction. The decoder neural network $p_{\boldsymbol{\theta}}$ receives the extracted features $\boldsymbol{Z}^{(i)}$ and makes a segmentation prediction $\hat{x}^{(i,c)}_{i_1,i_2,i_3} \in [0,1]$ for each voxel with indices $i_1 \in \{1,\ldots,N_1\}, i_2 \in \{1,\ldots,N_2\}, i_3 \in \{1,\ldots,N_3\}$. $\hat{x}^{(i,c)}_{i_1,i_2,i_3}$ represents the probability that the voxel (i_1,i_2,i_3) of image i belongs to class $c \in \{1,\ldots,C\}$. We denote the 3D segmentation prediction for each class via the tensor $\hat{\boldsymbol{X}}^{(i,c)} \in [0,1]^{N_1 \times N_2 \times N_3}$, in which the voxel (i_1,i_2,i_3) takes the class probability value $\hat{x}^{(i,c)}_{i_1,i_2,i_3}$.

2.1 Comparative Unsupervised Pre-training

Our aim is to effectively initialize the encoder-decoder architecture parameters $\boldsymbol{\phi} \in \mathbb{R}^{d_g}$ and $\boldsymbol{\theta} \in \mathbb{R}^{d_p}$ prior to supervised fine-tuning on ground-truth segments. To do so, we propose to pre-train $g_{\boldsymbol{\phi}}$ and $p_{\boldsymbol{\theta}}$ to impose a global structure that is unique to 3D medical images. The fundamental idea behind the global structure is that 2D slices corresponding to the same anatomical regions across different images are *more similar* to each other than other slices corresponding to different anatomical regions.

To group 2D slices based on anatomical regions along each axis, we partition each of the 3 dimensions N_1, N_2, and N_3 into P partitions, employing the same partitioning across all M images. For each pair of images $i,j \in \{1,\ldots,M\}$ for $i \neq j$, we impose the similarity of the same anatomical regions via the similarity of the corresponding 2D slices in each partition $p \in \{1,\ldots,P\}$. We denote a 2D slice sampled from the partition p in image $\boldsymbol{X}^{(i)}$ along the 3 axes by $\boldsymbol{X}^{(i)}_{p_1}$, $\boldsymbol{X}^{(i)}_{p_2}$, and $\boldsymbol{X}^{(i)}_{p_3}$, respectively. As a result, for each image pair $i,j \in \{1,\ldots,M\}$ and partition $p \in \{1,\ldots,P\}$, $\boldsymbol{X}^{(i)}_{p_1}$ and $\boldsymbol{X}^{(j)}_{p_1}$, $\boldsymbol{X}^{(i)}_{p_2}$ and $\boldsymbol{X}^{(j)}_{p_2}$, and $\boldsymbol{X}^{(i)}_{p_3}$ and $\boldsymbol{X}^{(j)}_{p_3}$ are similar 2D slice pairs.

Learning similarities via a DNN architecture requires the existence of dissimilar samples along with similar ones [33]. For instance, training the decoder $p_{\boldsymbol{\theta}}$ to simply make $\boldsymbol{X}^{(i)}_{p_1}$ as similar as possible to $\boldsymbol{X}^{(j)}_{p_1}$ would lead to a trivial solution, as the difference between $\boldsymbol{X}^{(i)}_{p_1}$ and $\boldsymbol{X}^{(j)}_{p_1}$ can be made zero by setting $p_{\boldsymbol{\theta}}$ output to an arbitrary constant. Thus, we focus on *comparative* pre-training on triplets of 2D slices, where each triplet comprises a reference slice, a similar slice from the same anatomical region as the reference, and a dissimilar slice from a different anatomical region. Without loss of generality, we focus on the first axis out of the 3 axes, as our approach applies in the same fashion to other 2 axes. As a result, for each image pair $i,j \in \{1,\ldots,M\}$ and partition pair $p,\tilde{p} \in \{1,\ldots,P\}$ for $p \neq \tilde{p}$, $\boldsymbol{X}^{(i)}_{p_1}$, $\boldsymbol{X}^{(j)}_{p_1}$ and $\boldsymbol{X}^{(i)}_{\tilde{p}_1}$ form a comparative slice triplet, where $\boldsymbol{X}^{(i)}_{p_1}$ from partition p is more similar to $\boldsymbol{X}^{(j)}_{p_1}$ than $\boldsymbol{X}^{(i)}_{\tilde{p}_1}$ from a different partition $\tilde{p}$.

Comparative Pre-training Objective. We aim to pre-train the encoder-decoder DNN architecture on the comparative slice triplets to impose global structure. Thus, for each slice triplet $\boldsymbol{X}^{(i)}_{p_1}$, $\boldsymbol{X}^{(j)}_{p_1}$ and $\boldsymbol{X}^{(i)}_{\tilde{p}_1}$, and segment class $c \in \{1,\ldots,C\}$, we impose the comparative structure on the corresponding decoder predictions $\hat{\boldsymbol{X}}^{(i,c)}_{p_1}$, $\hat{\boldsymbol{X}}^{(j,c)}_{p_1}$ and $\hat{\boldsymbol{X}}^{(i,c)}_{\tilde{p}_1}$, as illustrated in Figure 1.

To this end, we employ a prevalent probabilistic comparison model called the Bradley-Terry model [6]. The Bradley-Terry model asserts that every compared sample is parametrized by a positive deterministic score s. As our aim is to compare 2D slice pairs in terms of structural similarity, we employ the score $s(\hat{\boldsymbol{X}}^{(i,c)}_{p_1}, \hat{\boldsymbol{X}}^{(j,c)}_{p_1})$ as the similarity of predictions $\hat{\boldsymbol{X}}^{(i,c)}_{p_1}$ and $\hat{\boldsymbol{X}}^{(j,c)}_{p_1}$. We use the cosine similarity as a common similarity measure in multiple dimensions [66]. Then, under the Bradley-Terry model, all comparison events are independent of each other, where the probability that $\hat{\boldsymbol{X}}^{(i,c)}_{p_1}$ is more similar to $\hat{\boldsymbol{X}}^{(j,c)}_{p_1}$ than $\hat{\boldsymbol{X}}^{(i,c)}_{\tilde{p}_1}$ is given by:

$$\frac{s(\hat{\boldsymbol{X}}^{(i,c)}_{p_1}, \hat{\boldsymbol{X}}^{(j,c)}_{p_1})}{s(\hat{\boldsymbol{X}}^{(i,c)}_{p_1}, \hat{\boldsymbol{X}}^{(j,c)}_{p_1}) + s(\hat{\boldsymbol{X}}^{(i,c)}_{p_1}, \hat{\boldsymbol{X}}^{(i,c)}_{\tilde{p}_1})}, \quad \text{for all } i,j \in \{1,\ldots,M\};\ p,\tilde{p} \in \{1,\ldots,P\}. \tag{1}$$

We propose to pre-train the encoder-decoder DNN architecture on all slice triplets, in which each pair of 2D slices sampled from the same partition is more similar to each other than a pair of 2D slices sampled from different partitions. To do so, we employ maximum likelihood estimation (MLE) under the Bradley-Terry model governed Eq. (1). MLE is equivalent to the well-known cross-entropy loss, applied in this setting on the difference between

scores $s(\hat{\boldsymbol{X}}_{p_1}^{(i,c)}, \hat{\boldsymbol{X}}_{p_1}^{(j,c)})$ and $s(\hat{\boldsymbol{X}}_{p_1}^{(i,c)}, \hat{\boldsymbol{X}}_{\tilde{p}_1}^{(i,c)})$ [78]. The resulting DNN pre-training objective is given by:

$$\min_{\boldsymbol{\phi}\in\mathbb{R}^{d_g},\boldsymbol{\theta}\in\mathbb{R}^{d_p}} \sum_{c\in\{1,\ldots,C\}} \sum_{i,j\in\{1,\ldots,M\}} \sum_{p,\tilde{p}\in\{1,\ldots,P\}} \log\left(1+e^{s(\hat{\boldsymbol{X}}_{p_1}^{(i,c)},\hat{\boldsymbol{X}}_{\tilde{p}_1}^{(i,c)})-s(\hat{\boldsymbol{X}}_{p_1}^{(i,c)},\hat{\boldsymbol{X}}_{p_1}^{(j,c)})}\right). \tag{2}$$

Advantage Over Contrastive Pre-training. Our comparative pre-training strategy differs from contrastive pre-training [8] by using fine-grained global structure information via slice triplets, each comprising a reference slice, a slice from the same anatomical region that is similar to the reference, and a slice from a different anatomical region that is dissimilar to the reference. In contrast, contrastive pre-training asserts only the similarity of corresponding 2D slice pairs across different images, which is normalized by the pair's dissimilarity to all other slices from all other anatomical regions. The resulting contrastive pre-training objective is given by:

$$\min_{\boldsymbol{\phi}\in\mathbb{R}^{d_g},\boldsymbol{\theta}\in\mathbb{R}^{d_p}} \sum_{c\in\{1,\ldots,C\}} \sum_{i,j\in\{1,\ldots,M\}} \sum_{p\in\{1,\ldots,P\}} -\log \frac{e^{s(\hat{\boldsymbol{X}}_{p_1}^{(i,c)},\hat{\boldsymbol{X}}_{p_1}^{(j,c)})}}{e^{s(\hat{\boldsymbol{X}}_{p_1}^{(i,c)},\hat{\boldsymbol{X}}_{p_1}^{(j,c)})}+\sum_{\tilde{p}\in\{1,\ldots,P\}\setminus p} e^{s(\hat{\boldsymbol{X}}_{p_1}^{(i,c)},\hat{\boldsymbol{X}}_{\tilde{p}_1}^{(i,c)})}}. \tag{3}$$

Eq. (3) *does not* scale well with the number of partitions P in terms of computational complexity. The existence of 2D slice samples from all other partitions in the denominator requires evaluation and back-propagation over all such samples in each training step. In contrast, 2D slice samples from other partitions appear over *a sum rather than a division* in each training step of the novel comparative strategy (2). This allows the evaluations and parameter updates over different partition samples to be treated the same way as different slice samples, and thus enables significant acceleration through efficient batch-based training [45, 51]. Overall, comparative pre-training is advantageous over contrastive pre-training by: (i) revealing more fine-grained global structure information, and (ii) leading to significantly faster training by employing reference, similar, and dissimilar triplets rather than dissimilarity to all other slices from different anatomical regions. We experimentally demonstrate these benefits for the downstream segmentation task in Section 3.

3 Experiments

Datasets. We evaluate our approach over two benchmark datasets provided by the Medical Image Segmentation Decathlon [65]: Task 1 for brain tumour segmentation on Magnetic Resonance Imaging (MRI) and Task 9 for spleen segmentation on Computed Tomography (CT). Task 1 contains $M = 484$ 3D images with four modalities (FLAIR, T1w, T1gd, T2w) and three ground-truth segments. Following Hatamizadeh et al. [34], we map these to $C = 3$ classes called Tumor core (TC), Whole tumor (WT), and Enhancing tumor (ET), and perform segmentation for each class vs. the background. TC comprises non-enhancing tumor core and enhancing tumor, and WT comprises all three ground-truth segments. Task 9 contains $M = 41$ 3D CT images with $C = 1$ segment class indicating spleen regions, resulting in a binary segmentation task. We resample all images to ensure 1.0 mm isotropic spacing between voxels, and further normalize voxel intensities within the range $[0, 1]$ via z-score normalization.

Experiment Setup. We partition each dataset into training and test sets, allocating 80% of the images for training, and the remaining 20% for testing. We augment the training set by extracting 3D cubical crops from each image, transformed via random flipping, scaling, and intensity shifting. We choose the 3D crop size for Task 1 as 128 and Task 9 as 96. We implement [1] our approach on the UNET transformer architecture (UNETR); we refer the reader to Appendix B for details on UNETR.

Unsupervised Pre-training. We pre-train UNETR on the training set images via our unsupervised comparative objective (2), employing AdamW optimization [50] with a learning rate of 0.0001. We pre-train until the convergence of objective values, with a maximum of $10,000$ training steps. Convergence is declared if the current objective value does not differ significantly from the running objective average over the latest 10 steps, for which the criteria is remaining within 0.01 times the absolute value of the running average. Following Chaitanya et al. [8], we select the number of partitions for each axis as $P = 4$, and pre-train encoder and decoder stagewise to aid learning for both latent feature extraction and reconstruction. To do so, the encoder architecture $g_{\boldsymbol{\phi}}$ is pre-trained first using the objective (2), where $\hat{\boldsymbol{X}}^{(i,c)}$ is replaced by $\boldsymbol{Z}^{(i)}$ for $i \in \{1, \ldots, M\}$ to impose global structure via the similarity of latent features extracted from the same anatomical regions. After encoder pre-training, the overall encoder-decoder DNN architecture is pre-trained using the objective (2) to impose global structure via the similarity of reconstructed 2D slices from the same anatomical regions. Finally, to generalize pre-training to all 3 image axes, the DNN architecture parameters $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are trained stagewise via (2), using p_1, p_2 and p_3 sequentially at each training step. To illustrate the acceleration benefit of our comparative objective, we additionally pre-train UNETR using the contrastive objective (3), using the same experiment setup as comparative pre-training.

[1] Our code is publicly available at `https://github.com/ilkyyldz95/3DmedicalImageSegmentation`

Table 1: Pre-training time (in hours) and segmentation performance of our comparative semi-supervised (Comp. Semi-sup.) approach vs. contrastive semi-supervised (Cont. Semi-sup.) learning and supervised learning with no pre-training over the Task 1 and Task 9 datasets. For Task 1, we repeat fine-tuning experiments by subsampling the segment labels in the training set, with number of labeled images given in column 2. Lower HSD, higher Dice, and higher F1 indicate better segmentation.

Dataset	Labels	Method	Pre-train Time ↓	Segmentation Performance on Test Set								
				Dice ↑			F1 ↑			HSD ↓		
Task 9	32	**Comp. Semi-sup.**	**0.3**	**0.92**			**0.99**			**241.2**		
		Cont. Semi-sup.	5.27	0.92			0.99			266.15		
		Supervised	N/A	0.86			0.99			284.43		
				TC	WT	ET	TC	WT	ET	TC	WT	ET
Task 1	2	**Comp. Semi-sup.**	**16.38**	**0.46**	**0.73**	**0.56**	0.52	**0.76**	**0.68**	163.55	97.83	99.44
		Cont. Semi-sup.	116.66	0.46	0.68	0.51	**0.53**	0.71	0.63	188	175.48	175.45
		Supervised	N/A	0.39	0.62	0.45	0.47	0.7	0.6	**132.05**	**83.81**	**91.77**
Task 1	5	**Comp. Semi-sup.**	**16.38**	**0.6**	**0.78**	**0.65**	**0.67**	**0.81**	**0.79**	**82.08**	**73.6**	71.4
		Cont. Semi-sup.	116.66	0.58	0.76	0.6	0.64	0.79	0.76	92.26	82.02	**69.78**
		Supervised	N/A	0.6	0.77	0.63	0.66	0.81	0.77	113.72	138.32	80.49
Task 1	387	**Comp. Semi-sup.**	**16.38**	**0.78**	**0.88**	**0.73**	0.81	**0.89**	**0.85**	35.26	42.62	38.56
		Cont. Semi-sup.	116.66	0.76	0.87	0.72	0.8	0.89	0.85	42.39	40.33	30.79
		Supervised	N/A	0.77	0.88	0.73	**0.82**	0.89	0.85	**32.68**	**33.86**	**23.59**

Semi-supervised Segmentation. Following pre-training, we fine-tune UNETR with supervised learning from ground-truth segments in the training set, additionally augmenting the training set via crops extracted separately from segment vs. background regions. Following Hatamizadeh et al. [34], we perform segmentation training via dice loss combined with cross-entropy loss on each voxel, using AdamW optimization with a learning rate of 0.0001 for $25,000$ steps. We perform semi-supervised segmentation by supervised fine-tuning following three pre-training strategies: (i) random initialization, i.e., no pre-training, (ii) unsupervised pre-training via contrastive strategy (3), and (iii) unsupervised pre-training via our novel comparative strategy (2). To further illustrate the benefit of semi-supervised segmentation in label-deficient settings, we subsample the segmentation-labeled training set of the larger dataset of Task 1 and repeat the fine-tuning.

Evaluation Metrics. We report the elapsed time until the convergence of pre-training in hours. Following unsupervised pre-training and supervised fine-tuning on the training set, we report the segmentation prediction performance over the test set. We report benchmark evaluation metrics based on the (i) overlap of predicted and ground-truth segments, including Dice similarity score and F1 score, as well as (ii) maximal Euclidean distance between the voxels in the predicted vs. ground-truth segments, including Hausdorff surface distance (HSD) [80].

Pre-training Speed Results. Column 4 in Table 1 shows the pre-training time of our comparative semi-supervised (Comp. Semi-sup.) approach vs. contrastive semi-supervised (Cont. Semi-sup.) learning over Task 1 and Task 9. Our novel comparative pre-training objective (2) leads to *up to* 17 *times faster training* than the state-of-the-art contrastive pre-training objective (3). This observation validates the hypothesized speed gain due to employing comparative slice triplets rather than dissimilarity to all other slices from different anatomical regions. Note that the pre-training time resulting from our approach *does not* bring in significant additional training time complexity prior to supervised fine-tuning, which takes on average 8 hours using an NVIDIA V100 GPU.

Segmentation Results. Columns 5-7 in Table 1 show the segmentation prediction performance of Comp. Semi-sup. vs. Cont. Semi-sup. and supervised learning with no pre-training over Task 1 and Task 9. For the larger dataset of Task 1, we subsample the labeled training set and repeat the fine-tuning experiments. First and foremost, as pre-training captures global structure via the similarity of corresponding anatomical regions, segmentation predictions following both contrastive and comparative pre-training outperform the predictions following no pre-training with respect to most metrics. The performance improvement from *semi-supervised learning is the most significant for label-deficient settings*: comparative semi-supervised learning outperforms supervised learning by up to 11% Dice and 8% F1 over Task 1 with 2 training labels, and up to 47% HSD over Task 1 with 5 training labels, reducing the gap with learning from 75 times more labels that exist in the full training set. Moreover, our comparative pre-training strategy leads to *consistently better segmentation learning than contrastive pre-training*, by up to 5% Dice, 5% F1, and 47% HSD over Task 1 with 2 labeled training images. This observation further validates the hypothesized fine-grained information and prediction performance gain from comparative pre-training.

Visual Assessment of Segmentation. Figure 2 demonstrates example 2D slices from the test set of Task 1 with

ground-truth segments vs. predictions from supervised segmentation with no pre-training vs. our comparative semi-supervised approach. Our approach leads to significantly more visually similar predictions to the ground-truth segments compared to the supervised baseline. Over the images in rows 1-4, the incorrectly overestimated ET and TC segments by the supervised baseline are much larger than the ones predicted by our approach. Meanwhile, both methods underestimate the ET and TC segments on the image in row 5, while our approach is significantly more similar to the ground-truth than the supervised baseline.

4 Conclusion

We provide an efficient semi-supervised DNN method for automated 3D medical image segmentation. By incorporating the novel comparative pre-training strategy, we significantly accelerate DNN training over existing semi-supervised methods. Moreover, by focusing on semi-supervised rather than supervised segmentation, attaining accurate segmentation predictions via our method does not require large datasets of labeled images. More efficient, accurate, and easy to interpret automated segmentation via our approach can significantly help clinicians save time and effort in diagnosing disease, evaluating prognosis, and planning treatments.

Bibliography

[1] (2019). Unsupervised domain adaptation for medical imaging segmentation with self-ensembling. *NeuroImage*, 194:1–11.

[2] Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., and Asari, V. K. (2019). Recurrent residual U-Net for medical image segmentation. *Journal of Medical Imaging*, 6(1):014006.

[3] Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

[4] Bai, W., Suzuki, H., Qin, C., Tarroni, G., Oktay, O., Matthews, P. M., and Rueckert, D. (2018). Recurrent neural networks for aortic image sequence segmentation with sparse annotations. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 586–594. Springer.

[5] Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., Cetin, I., Lekadir, K., Camara, O., Ballester, M. A. G., et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved? *IEEE Transactions on Medical Imaging*, 37(11):2514–2525.

[6] Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

[7] Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., and Wang, M. (2021). Swin-U-net: U-net-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*.

[8] Chaitanya, K., Erdil, E., Karani, N., and Konukoglu, E. (2020). Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33.

[9] Chaudhari, S., Mithal, V., Polatkan, G., and Ramanath, R. (2019). An attentive survey of attention models. *arXiv preprint arXiv:1904.02874*.

[10] Chen, H., Dou, Q., Yu, L., and Heng, P.-A. (2016a). Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation. *arXiv preprint arXiv:1608.05895*.

[11] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., and Zhou, Y. (2021). TransU-net: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

[12] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*.

[13] Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR.

[14] Chen, W., Smith, R., Ji, S.-Y., Ward, K. R., and Najarian, K. (2009). Automated ventricular systems segmentation in brain CT images by combining low-level segmentation and high-level template matching. *BMC Medical Informatics and Decision Making*, 9(1):1–14.

[15] Chen, X., Xu, L., Yang, Y., and Egger, J. (2016b). A semi-automatic computer-aided method for surgical template design. *Scientific Reports*, 6(1):1–18.

[16] Chen, X., Zhang, R., and Yan, P. (2019). Feature fusion encoder decoder network for automatic liver lesion segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 430–433. IEEE.

[17] Cheng, J., Liu, J., Xu, Y., Yin, F., Wong, D. W. K., Tan, N.-M., Tao, D., Cheng, C.-Y., Aung, T., and Wong, T. Y. (2013). Superpixel classification based optic disc and optic cup segmentation for glaucoma screening. *IEEE Transactions on Medical Imaging*, 32(6):1019–1032.

[18] Cherukuri, V., Ssenyonga, P., Warf, B. C., Kulkarni, A. V., Monga, V., and Schiff, S. J. (2017). Learning based segmentation of CT brain images: application to postoperative hydrocephalic scans. *IEEE Transactions on Biomedical Engineering*, 65(8):1871–1884.

[19] Chrástek, R., Wolf, M., Donath, K., Niemann, H., Paulus, D., Hothorn, T., Lausen, B., Lämmer, R., Mardin, C. Y., and Michelson, G. (2005). Automated segmentation of the optic nerve head for diagnosis of glaucoma. *Medical Image Analysis*, 9(4):297–314.

[20] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., and Ronneberger, O. (2016). 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.

[21] Dong, H., Yang, G., Liu, F., Mo, Y., and Guo, Y. (2017). Automatic brain tumor detection and segmentation using U-Net based fully convolutional networks. In *Annual Conference on Medical Image Understanding and Analysis*, pages 506–517. Springer.

[22] Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van Der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341–355.

[23] Formisano, E. and Goebel, R. (2003). Tracking cognitive processes with functional MRI mental chronometry. *Current opinion in neurobiology*, 13(2):174–181.

[24] Fortunati, V., Verhaart, R. F., van der Lijn, F., Niessen, W. J., Veenland, J. F., Paulides, M. M., and van Walsum, T. (2013). Tissue segmentation of head and neck CT images for treatment planning: a multiatlas approach combined with intensity modeling. *Medical Physics*, 40(7):071905.

[25] Fu, H., Cheng, J., Xu, Y., Wong, D. W. K., Liu, J., and Cao, X. (2018). Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Transactions on Medical Imaging*, 37(7):1597–1605.

[26] Gao, Y., Phillips, J. M., Zheng, Y., Min, R., Fletcher, P. T., and Gerig, G. (2018). Fully convolutional structured LSTM networks for joint 4d medical image segmentation. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1104–1108. IEEE.

[27] Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I. W., Sanchez, C. I., Litjens, G., de Leeuw, F.-E., van Ginneken, B., Marchiori, E., and Platel, B. (2017). Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *Scientific Reports*, 7(1):1–12.

[28] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT press.

[29] Gu, R., Wang, G., Song, T., Huang, R., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., and Zhang, S. (2020). Ca-net: Comprehensive attention convolutional neural networks for explainable medical image segmentation. *IEEE Transactions on Medical Imaging*, 40(2):699–711.

[30] Gu, Z., Cheng, J., Fu, H., Zhou, K., Hao, H., Zhao, Y., Zhang, T., Gao, S., and Liu, J. (2019). Ce-net: Context encoder network for 2d medical image segmentation. *IEEE Transactions on Medical Imaging*, 38(10):2281–2292.

[31] Guan, S., Khan, A. A., Sikdar, S., and Chitnis, P. V. (2019). Fully dense U-net for 2-d sparse photoacoustic tomography artifact removal. *IEEE Journal of Biomedical and Health Informatics*, 24(2):568–576.

[32] Guibas, J. T., Virdi, T. S., and Li, P. S. (2017). Synthetic medical images from dual generative adversarial networks. *arXiv preprint arXiv:1709.01872*.

[33] Hadsell, R., Chopra, S., and LeCun, Y. (2006). Dimensionality reduction by learning an invariant mapping. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742. IEEE.

[34] Hatamizadeh, A., Yang, D., Roth, H., and Xu, D. (2021). U-netr: Transformers for 3D medical image segmentation. *arXiv preprint arXiv:2103.10504*.

[35] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

[36] Hesamian, M. H., Jia, W., He, X., and Kennedy, P. (2019). Deep learning techniques for medical image segmentation: achievements and challenges. *Journal of digital imaging*, 32(4):582–596.

[37] Hu, J., Shen, L., and Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7132–7141.

[38] Kaul, C., Manandhar, S., and Pears, N. (2019). Focusnet: An attention-based fully convolutional network for medical image segmentation. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 455–458. IEEE.

[39] Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., and Shah, M. (2021). Transformers in vision: A survey. *arXiv preprint arXiv:2101.01169*.

[40] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105.

[41] Lalonde, M., Beaulieu, M., and Gagnon, L. (2001). Fast and robust optic disc detection using pyramidal decomposition and hausdorff-based template matching. *IEEE Transactions on Medical Imaging*, 20(11):1193–1200.

[42] LeCun, Y., Bengio, Y., Hinton, G., et al. (2015). Deep learning. 521 (7553), 436-444. *Nature*.

[43] Lei, T., Wang, R., Wan, Y., Zhang, B., Meng, H., and Nandi, A. K. (2020). Medical image segmentation using deep learning: a survey. *arXiv preprint arXiv:2009.13120*.

[44] Li, C., Wang, X., Eberl, S., Fulham, M., Yin, Y., Chen, J., and Feng, D. D. (2013). A likelihood and local constraint level set model for liver tumor segmentation from CT volumes. *IEEE Transactions on Biomedical Engineering*, 60(10):2967–2977.

[45] Li, M., Zhang, T., Chen, Y., and Smola, A. J. (2014). Efficient mini-batch training for stochastic optimization. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 661–670.

[46] Li, S., Fevens, T., and Krzyżak, A. (2004). A SVM-based framework for autonomous volumetric medical image segmentation using hierarchical and coupled level sets. In *International Congress Series*, volume 1268, pages 207–212. Elsevier.

[47] Li, W. et al. (2015). Automatic segmentation of liver tumor in CT images with deep convolutional neural networks. *Journal of Computer and Communications*, 3(11):146.

[48] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., Van Der Laak, J. A., Van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88.

[49] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

[50] Loshchilov, I. and Hutter, F. (2017). Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[51] Masters, D. and Luschi, C. (2018). Revisiting small batch training for deep neural networks. *arXiv preprint arXiv:1804.07612*.

[52] Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al. (2014). The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024.

[53] Mezer, A., Yeatman, J. D., Stikov, N., Kay, K. N., Cho, N.-J., Dougherty, R. F., Perry, M. L., Parvizi, J., Hua, L. H., Butts-Pauly, K., et al. (2013). Quantifying the local tissue volume and composition in individual brains with magnetic resonance imaging. *Nature medicine*, 19(12):1667–1672.

[54] Milletari, F., Navab, N., and Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D Vision (3DV)*, pages 565–571. IEEE.

[55] Norouzi, A., Rahim, M. S. M., Altameem, A., Saba, T., Rad, A. E., Rehman, A., and Uddin, M. (2014). Medical image segmentation methods, algorithms, and applications. *IETE Technical Review*, 31(3):199–213.

[56] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N. Y., Kainz, B., et al. (2018). Attention U-net: Learning where to look for the pancreas. *MIDL*.

[57] Onishi, Y., Teramoto, A., Tsujimoto, M., Tsukamoto, T., Saito, K., Toyama, H., Imaizumi, K., and Fujita, H. (2020). Multiplanar analysis for pulmonary nodule classification in CT images using deep convolutional neural network and generative adversarial networks. *International Journal of Computer Assisted Radiology and Surgery*, 15(1):173–178.

[58] Peng, J., Estrada, G., Pedersoli, M., and Desrosiers, C. (2020). Deep co-training for semi-supervised image segmentation. *Pattern Recognition*, 107:107269.

[59] Renard, F., Guedria, S., De Palma, N., and Vuillerme, N. (2020). Variability and reproducibility in deep learning for medical image segmentation. *Scientific Reports*, 10(1):1–16.

[60] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham. Springer International Publishing.

[61] Salehi, S. S. M., Erdogmus, D., and Gholipour, A. (2017). Tversky loss function for image segmentation using 3d fully convolutional deep networks. In *International Workshop on Machine Learning in Medical Imaging*, pages 379–387. Springer.

[62] Sharma, K., Rupprecht, C., Caroli, A., Aparicio, M. C., Remuzzi, A., Baust, M., and Navab, N. (2017). Automatic segmentation of kidneys using deep learning for total kidney volume quantification in autosomal dominant polycystic kidney disease. *Scientific Reports*, 7(1):1–10.

[63] Sharma, N. and Aggarwal, L. M. (2010). Automated medical image segmentation techniques. *Journal of Medical Physics/Association of Medical Physicists of India*, 35(1):3.

[64] Silveira, M., Nascimento, J. C., Marques, J. S., Marçal, A. R., Mendonça, T., Yamauchi, S., Maeda, J., and Rozeira, J. (2009). Comparison of segmentation methods for melanoma diagnosis in dermoscopy images. *IEEE Journal of Selected Topics in Signal Processing*, 3(1):35–45.

[65] 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. (2019). A large annotated medical image dataset for the development and evaluation of segmentation algorithms. *arXiv preprint arXiv:1902.09063*.

[66] Singhal, A. et al. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4):35–43.

[67] Sirinukunwattana, K., Pluim, J. P., Chen, H., Qi, X., Heng, P.-A., Guo, Y. B., Wang, L. Y., Matuszewski, B. J., Bruni, E., Sanchez, U., et al. (2017). Gland segmentation in colon histology images: The glas challenge contest. *Medical Image Analysis*, 35:489–502.

[68] Song, T.-H., Sanchez, V., EIDaly, H., and Rajpoot, N. M. (2017). Dual-channel active contour model for megakaryocytic cell segmentation in bone marrow trephine histology images. *IEEE Transactions on Biomedical Engineering*, 64(12):2913–2923.

[69] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. (2017). Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer.

[70] Trebeschi, S., van Griethuysen, J. J., Lambregts, D. M., Lahaye, M. J., Parmar, C., Bakers, F. C., Peters, N. H., Beets-Tan, R. G., and Aerts, H. J. (2017). Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Scientific Reports*, 7(1):1–9.

[71] Tsai, A., Yezzi, A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, W. E., and Willsky, A. (2003). A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Transactions on Medical Imaging*, 22(2):137–154.

[72] Tu, Z., Narr, K. L., Dollár, P., Dinov, I., Thompson, P. M., and Toga, A. W. (2008). Brain anatomical structure segmentation by hybrid discriminative/generative models. *IEEE Transactions on Medical Imaging*, 27(4):495–508.

[73] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

[74] Vivanti, R., Ephrat, A., Joskowicz, L., Karaaslan, O., Lev-Cohain, N., and Sosna, J. (2015). Automatic liver tumor segmentation in follow-up CT studies using convolutional neural networks. In *Proc. Patch-Based Methods in Medical Image Processing Workshop*, volume 2, page 2.

[75] Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., Dong, D., Gevaert, O., and Tian, J. (2017). Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. *Medical Image Analysis*, 40:172–183.

[76] Withey, D. and Koles, Z. (2008). A review of medical image segmentation: Methods and available software.

[77] Xiao, X., Lian, S., Luo, Z., and Li, S. (2018). Weighted Res-U-net for high-quality retina vessel segmentation. In *International Conference on Information Technology in Medicine and Education (ITME)*, pages 327–331. IEEE.

[78] Yıldız, İ., Tian, P., Dy, J., Erdoğmuş, D., Brown, J., Kalpathy-Cramer, J., Ostmo, S., Campbell, J. P., Chiang, M. F., and Ioannidis, S. (2019). Classification and comparison via neural networks. *Neural Networks*, 118:65–80.

[79] You, C., Zhao, R., Staib, L., and Duncan, J. S. (2021a). Momentum contrastive voxel-wise representation learning for semi-supervised volumetric medical image segmentation. *arXiv preprint arXiv:2105.07059*.

[80] You, J., Philip, L., Tsang, A. C., Tsui, E. L., Woo, P. P., Lui, C. S., Leung, G. K., Mahboobani, N., Chu, C.-y., Chong, W.-h., et al. (2021b). 3d dissimilar-siamese-u-net for hyperdense middle cerebral artery sign segmentation. *Computerized Medical Imaging and Graphics*, 90:101898.

[81] Yu-Qian, Z., Wei-Hua, G., Zhen-Cheng, C., Jing-Tian, T., and Ling-Yun, L. (2006). Medical images edge detection based on mathematical morphology. In *2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, pages 6492–6495. IEEE.

[82] Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D. P., and Chen, D. Z. (2017). Deep adversarial networks for biomedical image segmentation utilizing unannotated images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 408–416. Springer.

[83] Zhou, Y., Wang, Y., Tang, P., Bai, S., Shen, W., Fishman, E. K., and Yuille, A. L. (2018). Semi-supervised multi-organ segmentation via deep multi-planar co-training. *arXiv preprint arXiv:1804.02586*.

[84] Zhou, Z., Sodha, V., Siddiquee, M. M. R., Feng, R., Tajbakhsh, N., Gotway, M. B., and Liang, J. (2019). Models genesis: Generic autodidactic models for 3d medical image analysis. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 384–393. Springer.

[85] Zhuang, X. and Shen, J. (2016). Multi-scale patch and multi-modality atlases for whole heart segmentation of MRI. *Medical Image Analysis*, 31:77–87.

A Related Work

Medical image segmentation has various applications, including tissue volume estimation [53, 62], diagnosis [64, 19], pathology localization [27, 70], anatomical structure analysis [72, 22], treatment planning [24], and computer-aided surgery [15]. Manual execution of medical image segmentation is a highly laborsome and time-consuming task for clinical experts [63, 55, 59]. To this end, automated segmentation provides critical information by improving diagnosis accuracy and diagnostic decision explainability in computer-aided healthcare and smart medicine applications [63]. Popular medical image segmentation tasks include liver and liver-tumor [47, 74], brain and brain-tumor [52, 18], cardiac structure [5, 85], organ [65], optic disc [17, 25], cell [60, 68], and nodule [75, 57], focusing on X-ray, Computed Tomography (CT), Magnetic Resonance Imaging (MRI) and ultrasound modalities.

Early approaches to automated medical image segmentation often involved edge detection [81], template matching [41, 14], and statistical shape [71] and contour [46, 44] estimation. In the last decade, the rapid development of automated segmentation algorithms is strongly connected with the rising success and popularity of machine learning. Particularly, a specific field of machine learning and artificial neural networks, called *deep learning* [42], has outperformed traditional segmentation methods. A neural network with several hidden layers is termed as a deep neural network (DNN) [28]. Particularly, convolutional DNNs successfully extract complex hierarchical feature representations of images, and thus ceased the need for manual feature extraction in medical image analysis [40].

There is a vast literature on DNN methods in medical image segmentation, we refer the reader to the surveys by Litjens et al. [48], Hesamian et al. [36], Lei et al. [43] for an extended review. Particularly, encoder-decoder DNN architectures have become state-of-the-art for this task [49, 60]. The encoder extracts latent features from a given image by gradual dimension reduction, while the decoder makes a reconstruction from the features in the original image size that captures the segmentation prediction for each voxel, i.e., 3D pixel. Arguably the most widely employed encoder-decoder DNN architecture for medical image segmentation is UNET [60], where the encoder and decoder networks are symmetrical and connected with additional connections at different resolutions. This design successfully handles the noise and resolution challenges in medical images. Many successful extensions on UNET have been proposed, including 3D image segmentation with few labeled 2D slices [20], deeper architecture designs with dense [54, 31] and residual [10, 77] connections, atrous convolutions [12, 30], and recurrent networks [2, 26, 4].

Existing DNN architectures for segmentation are trained via two approaches: *supervised* and *semi-supervised*. Supervision implies the existence of a dataset of images labeled by experts with ground-truth segments, which are used to train a DNN architecture. Trained DNN architectures are used to predict the segments on images that are unseen in training and are validated with their ground-truth segments. Segmentation architectures are typically trained with traditional cross-entropy for voxel-level classification [49], dice loss that maximizes the overlap between a segmentation prediction and the corresponding ground-truth [54], and their extensions to better handle the imbalance between segmentation-related vs. irrelevant image regions [61, 69].

In recent years, *attention* modules have significantly improved predictions in many machine learning applications including segmentation [37, 56, 16, 9, 38, 29]. An attention module is trained to selectively assign different importance to the input image regions and accordingly improve the downstream prediction task. Supervised segmentation architectures have been united and improved with spatial [56] and channel [37, 16] attention modules, along with their mixtures [38, 29]. Particularly, transformer architectures have acquired the most focus due to their self-attention modules [73, 39]. The self-attention module enables the transformer to dynamically highlight the important features and learn long-range multi-scale dependencies. Transformers have very recently been applied to medical image segmentation and combined with the traditional UNET [34, 7, 11]. Chen et al. [11] and Cao et al. [7] focus on 2D segmentation. Chen et al. [11] incorporate a transformer layer at the bottleneck of a UNET, while Cao et al. [7] employ a full encoder-decoder transformer architecture symmetrical encoder and decoder networks with skip connections inspired by the traditional UNET. We focus on the UNET Tranformer (UNETR) architecture, as it is tailored for 3D medical image segmentation. UNETR applies a transformer encoder on patches extracted from medical images, followed by a traditional UNET decoder to make segmentation predictions. This approach has become state-of-the-art among supervised DNN methods, with segmentation predictions leading to Dice overlap scores of 72% in brain tumour MRI segmentation and 96% in spleen CT segmentation [65].

In contrast to supervised DNN methods, semi-supervised approaches require the availability of only a small or partially-labeled image dataset with ground-truth segments. This is a realistic setting, as manual segmentation is a highly laborsome and time-consuming task for clinical experts [63, 55, 59]. Segmentation in label-deficient settings have been tackled with data augmentation via image transformations such as intensity variations and cropping [67, 21], synthetic data augmentation via generative models [32], adversarial learning [58, 82], and transfer learning across different machine learning applications [83, 1, 84]. More recently, initializing a DNN architecture by unsupervised *contrastive pre-training* prior to training on ground-truth segments led to significantly better predictions

than traditional supervised learning over few labels [13]. This approach trains the architecture by imposing that different transformations of an image should have similar features and that these features should be dissimilar from those of a different image. Contrastive pre-training has been recently employed in medical image segmentation and significantly improved predictions in cardiac and prostate segmentation based on MRI by incorporating global and local structures that are unique to 3D medical images [8, 79]. Global structure is learned by pre-training a DNN architecture to enforce similarity of corresponding 2D slices across different 3D images and their dissimilarity to all other 2D slices. We differ from the above approaches by employing a novel *comparative* pre-training strategy to initialize a DNN architecture for semi-supervised medical image segmentation, which leads to significant acceleration compared to contrastive pre-training (c.f. Section 3).

B The UNETR Architecture

Transformer architectures have been very recently unified with the prominent encoder-decoder UNET architecture [60] for supervised medical image segmentation [34, 7, 11]. We focus on the UNET Transformer (UNETR) architecture, as it is tailored for 3D medical image segmentation. UNETR enables segmentation via successful adaptations to long distance dependencies in various resolutions and provide important image regions for successful segmentation predictions. This approach has become state-of-the-art among supervised DNN methods, with segmentation predictions leading to Dice overlap scores of 72% in brain tumour MRI segmentation and 96% in spleen CT segmentation [65].

UNETR contains a transformer architecture as the encoder neural network $g_{\boldsymbol{\phi}}$, which is explained in detailed below. The decoder neural network $p_{\boldsymbol{\theta}}$ is a traditional convolutional UNET decoder [60] that receives the encoder features $\boldsymbol{Z}^{(i)}$ and makes a voxel-wise segmentation prediction $\hat{\boldsymbol{X}}^{(i,c)} \in [0,1]^{N_1 \times N_2 \times N_3}$ for each segmentation class $c \in \{1, \ldots, C\}$. Encoder and decoder networks are connected via additional skip connections that feed the outputs at intermediate layers of the transformer encoder and sums the predictions at intermediate layers of the decoder; this design enables information propagation in long distances and various resolutions. Following the novel comparative DNN pre-training strategy explained in Section 2.1, the UNETR architecture is fine-tuned via supervised training on a set of 3D images labeled with ground-truth segments to learn how to segment any given image $\boldsymbol{X}^{(i)}$.

The loss function that trains UNETR on ground-truth segments is a combination of the cross-entropy loss that classifies each voxel as a target segment vs. background [28], and the dice loss that is devised specifically for medical image segmentation [54]. In medical images, the segment interest typically occupies a small region of the entire image, such as a tumor in a brain MRI [52]. Learning from an imbalanced distribution of segment vs. background voxels leads the training via cross-entropy loss to be strongly biased towards background. That is why, given a segmentation prediction $\hat{\boldsymbol{X}}^{(i,c)}$ for image i and segmentation class c, Milletari et al. [54] proposes the dice loss for supervised training that maximizes the overlap between a target segment region and the decoder network predictions:

$$\min_{\boldsymbol{\phi} \in \mathbb{R}^{d_g}, \boldsymbol{\theta} \in \mathbb{R}^{d_p}} -\sum_{c=1}^{C} \sum_{i=1}^{M} \frac{\sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} \sum_{i_3=1}^{N_3} \hat{x}_{i_1,i_2,i_3}^{(i,c)} \, y_{i_1,i_2,i_3}^{(i,c)}}{\sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} \sum_{i_3=1}^{N_3} {y_{i_1,i_2,i_3}^{(i,c)}}^2 + \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} \sum_{i_3=1}^{N_3} {\hat{x}_{i_1,i_2,i_3}^{(i,c)}}^2}. \tag{4}$$

$y_{i_1,i_2,i_3}^{(i,c)}$ denotes the ground-truth segment for voxel (i_1, i_2, i_3), which takes on the value 1 if this voxel belongs to class $c \in \{1, \ldots, C\}$ and 0 otherwise.

Transformer Encoder. Inspired by the most well-known application of transformers in sequence modeling, UNETR creates a vector sequence from each 3D image $\boldsymbol{X}^{(i)}$ by partitioning $\boldsymbol{X}^{(i)}$ into $V \times V \times V$-dimensional voxel patches and vectorizing each patch [34]. As a result, each patch takes the form $\boldsymbol{x}^{(i)} \in \mathbb{R}^{L \times V^3}$ for $L = \frac{N_1 \times N_2 \times N_3}{V^3}$ and $i \in \{1, \ldots, M\}$. The transformer encoder $g_{\boldsymbol{\phi}}$ receives $\boldsymbol{x}^{(i)}$ and projects it from V^3 dimensions to D dimensions via a trainable linear transformation. To preserve the spatial information of the extracted patches, each projected patch is summed with a positional encoding matrix that is also trainable. The resulting latent features extracted from each patch is denoted as $\boldsymbol{z}^{(i)} \in \mathbb{R}^{L \times D}$. Dimension reduction and positional encoding is followed by the recursive application of several transformer layers. Each transformer layer applies a self-attention (SA) module, layer normalization (Norm) [3], and a fully-connected neural network (FCNN) [28]. Formally, each latent feature is updated by a transformer layer via the following steps:

$$\begin{aligned} \boldsymbol{z}^{(i)} &\leftarrow \text{SA}\big(\text{Norm}(\boldsymbol{z}^{(i)})\big) + \boldsymbol{z}^{(i)} \\ \boldsymbol{z}^{(i)} &\leftarrow \text{FCNN}\big(\text{Norm}(\boldsymbol{z}^{(i)})\big) + \boldsymbol{z}^{(i)}. \end{aligned} \tag{5}$$

The summation of each latent feature $\boldsymbol{z}^{(i)}$ with its transformed version is termed as a skip connection that aids resolution challenges in medical images [60, 35], while layer normalization aids the convergence of gradient-based

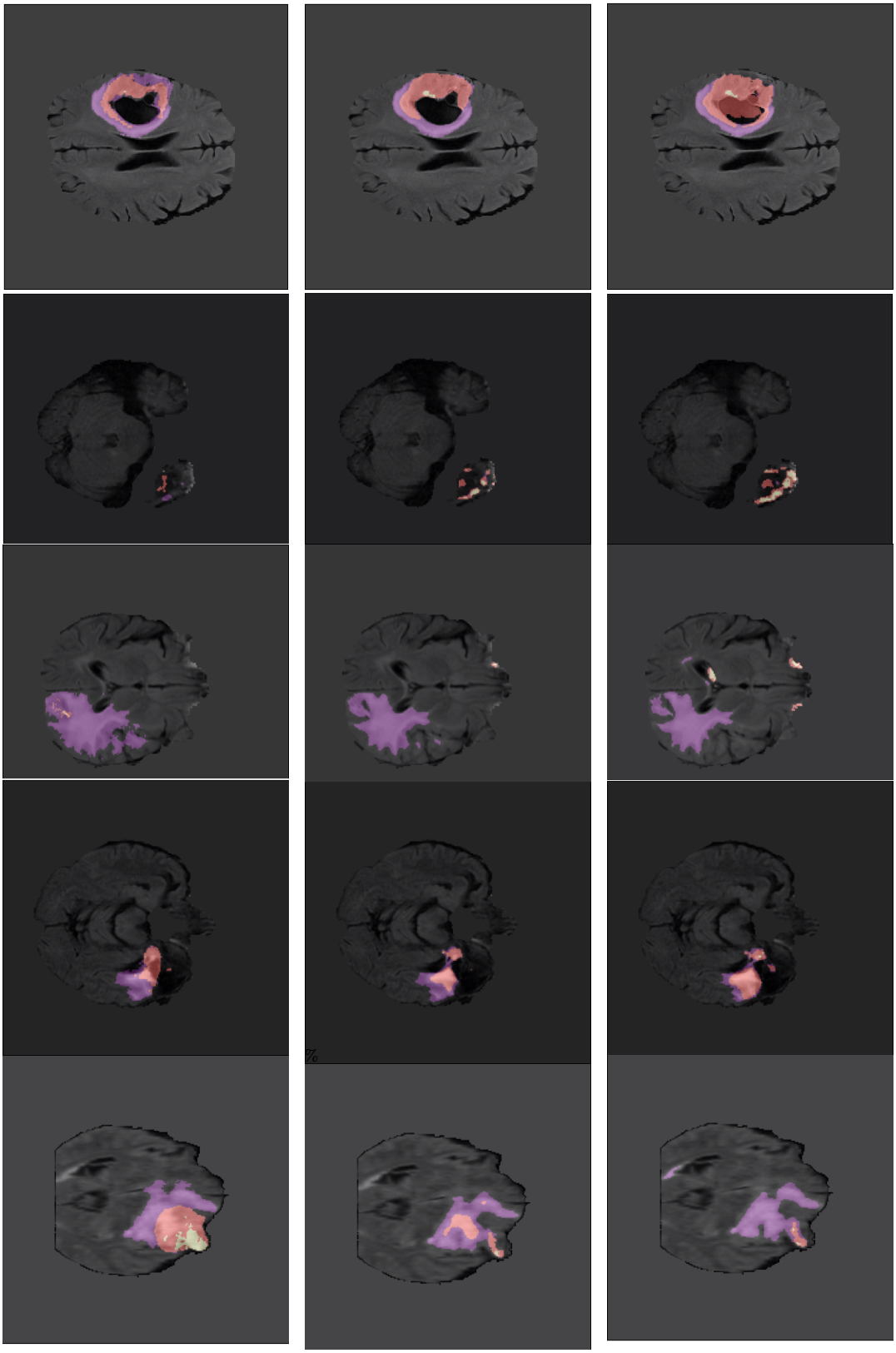
training over the segmentation loss (4). After the application of transformer layers, the transformed latent feature $\boldsymbol{z}^{(i)}$ is reshaped as $\boldsymbol{Z}^{(i)} \in \mathbb{R}^{\frac{N_1}{V} \times \frac{N_2}{V} \times \frac{N_3}{V}}$ before being fed to the decoder. Following Hatamizadeh et al. [34], we employ UNETR with voxel patch size $V = 16$, latent feature dimension $D = 768$, and 12 transformer layers, where FCNN comprises two fully-connected layers with GELU activations.

Self-attention Module. An SA module is designed to attend and assign different importance to all latent features learned by the previous layers of the encoder [73]. Particularly, SA contains trainable parameters that captures the similarity between input features via their query, key, and value representations. These representations are used to form an output via a weighted sum of the values, where the importance weight assigned to each value is computed by the similarity of the query with the corresponding key. Given a latent feature $\boldsymbol{z}^{(i)}$, an SA module first creates a query $\boldsymbol{z}_q^{(i)} \in \mathbb{R}^{L \times D_q}$, a key $\boldsymbol{z}_k^{(i)} \in \mathbb{R}^{L \times D_q}$, and a value $\boldsymbol{z}_v^{(i)} \in \mathbb{R}^{L \times D_v}$ by applying three different trainable linear transformations on $\boldsymbol{z}^{(i)}$. The SA output is then computed via a scaled dot-product:

$$\text{SA}(\boldsymbol{z}^{(i)}) = \text{softmax}\big(\frac{\boldsymbol{z}_q^{(i)} \boldsymbol{z}_k^{(i)\top}}{\sqrt{D_q}} \boldsymbol{z}_v^{(i)}\big), \tag{6}$$

where softmax is a commonly employed function for normalizing the output within the $[0, 1]$ range [28]. State-of-the-art transformer architectures [34, 7, 11] apply the SA module several times on the same latent features $\boldsymbol{z}^{(i)}$, concatenate the resulting outputs, and form the final attention module output via another trainable linear transformation after concatenation.

Figure 2: Example 2D slices from the test set of Task 1 with ground-truth segments vs. predictions from supervised segmentation with no pre-training vs. our comparative semi-supervised (Semi-sup.) approach. ET, TC, and WT classes are shaded in yellow, red, and purple, respectively.



Ground-truth

Semi-sup. prediction

Supervised prediction