Learning with Limited Annotations: A Survey on Deep Semi-Supervised Learning for Medical Image Segmentation

Rushi Jiao*, Yichi Zhang*, Le Ding, Rong Cai** and Jicong Zhang**

Abstract—Medical image segmentation is a fundamental and critical step in many image-guided clinical approaches. Recent success of deep learning-based segmentation methods usually relies on a large amount of labeled data, which is particularly difficult and costly to obtain especially in the medical imaging domain where only experts can provide reliable and accurate annotations. Semi-supervised learning has emerged as an appealing strategy and been widely applied to medical image segmentation tasks to train deep models with limited annotations. In this paper, we present a comprehensive review of recently proposed semi-supervised learning methods for medical image segmentation and summarized both the technical novelties and empirical results. Furthermore, we analyze and discuss the limitations and several unsolved problems of existing approaches. We hope this review could inspire the research community to explore solutions for this challenge and further promote the developments in medical image segmentation field.

Index Terms—Medical Image Segmentation, Semi-Supervised Learning, Convolutional Neural Network, Survey.

1 Introduction

Medical image segmentation aims to delineate the interested anatomical structures like organs and tumors from the original images by labeling each pixel into a certain class, which is a basic and important step for many clinical approaches like computer-aided diagnosis, treatment planning and radiation therapy [1], [2]. Accurate segmentation can provide reliable volumetric and shape information so as to assist in further clinical applications like disease diagnosis and quantitative analysis [3], [4], [5]. According to the word cloud of paper titles in the 24rd International Conference and Medical Image Computing and Computer Assisted Intervention [1] (MICCAI 2021) in Figure 1, we can observe that "segmentation" is one of the most active research topics and has the highest frequency in medical image analysis community.

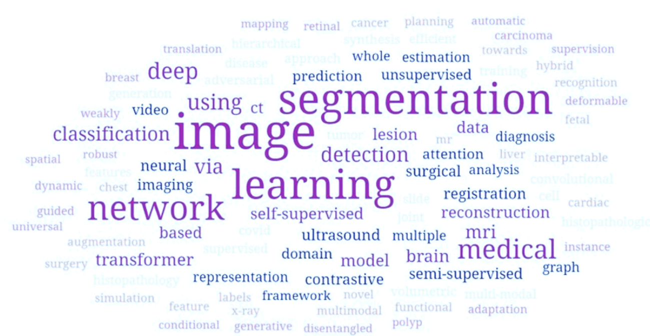


Fig. 1: Word cloud of paper titles in MICCAI 2021.

Since the introduction of U-Net [6], [7] for medical image segmentation in 2015, many variants of encoder-decoder architecture have been proposed to improve it by re-designing skip connections [8], incorporating residual/dense convolution blocks [9], [10], attention mechanisms [11], [12], etc. Isensee *et al.* [13] proposed nnU-Net to automatically adapt training strategies and network architectures to a given dataset for medical image segmentation based on encoder-decoder architecture. Inspired by recent success of transformer architectures in the field of natural language processing, many transformer-based methods have been proposed and applied for medical image segmentation [14], [15]. Although these architectural advancements have shown encouraging results and achieved state-of-the-art performances in many medical image segmentation tasks [16], these methods still require relatively large amount of high-quality annotated data for training, more than ever. However, it is impractical to obtain large-scale carefully-labeled datasets to train segmentation models, particularly for medical imaging where it is hard and expensive to obtain well-annotated data where only experts can provide reliable and accurate annotations [17]. Besides, many commonly used medical images like computed tomography (CT) and magnetic resonance imaging (MRI) scans are 3D volumetric data, which further increase the burden of manual annotation compared with 2D images where experts need to delineate the object from the volume slice by slice [18].

- *Rushi Jiao is with School of Biological Science and Medical Engineering, Beihang University, Beijing, China, and with School of Engineering Medicine, Beihang University, Beijing, China.*
- *Yichi Zhang, Le Ding are with School of Biological Science and Medical Engineering, Beihang University, Beijing, China.*
- *Rong Cai is with School of Engineering Medicine, Beihang University, Beijing, China, and Key Laboratory of Biomechanics and Mechanobiology (Beihang University), Ministry of Education.*
- *Jicong Zhang is with School of Biological Science and Medical Engineering, Beihang University, Beijing, China, and with Hefei Innovation Research Institute, Beihang University, Hefei, China.*

Current version of manuscript finished on July 27, 2022.
** Contribute equally to this work. ** Corresponding author.*

1. http://miccai2021.org

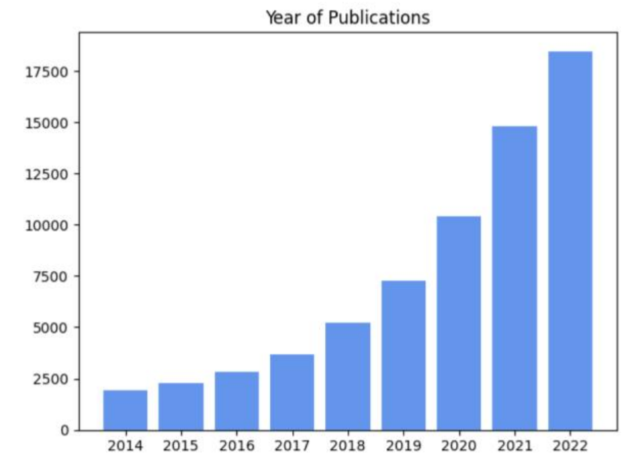


Fig. 2: Statistics of papers retrieved from Google Scholar on semi-supervised medical image segmentation. The data for the year 2022 has been extrapolated from the papers till June 2022.

To ease the manual labeling burden in response to these challenges, significant efforts have been devoted to annotation-efficient deep learning methods for medical image segmentation tasks by enlarging the training data through label generation [19], data augmentation [20], leveraging external related labeled datasets [21], and leveraging unlabeled data with semi-supervised learning. Among these approaches, semi-supervised segmentation is a more practical method by encouraging segmentation models to utilize unlabeled data which is much easier to acquire in conjunction with limited amount of labeled data for training, which has a high impact on real-world clinical applications. According to the statistics in Figure 2, semi-supervised medical image segmentation has obtained increasing attention from the medical imaging and computer vision community in recent years. However, without expert-examined annotations, it is still an open and challenging question on how to efficiently exploit useful information from these unlabeled data.

In this paper, we provide a comprehensive review of recent solutions for semi-supervised medical image segmentation and summarize both the technical novelties and empirical results. Furthermore, we analyze and discussed the limitations and several unsolved problems of existing approaches. We hope this review could inspire the research community to explore solutions for this challenge and further promote the developments in medical image segmentation field.

2 Preliminaries

2.1 Basic Formulation of Semi-Supervised Learning

Semi-supervised learning aims to utilize unlabeled data in conjunction with labeled data to train higher-performing segmentation models. To ease the description in the following sections, we formulate the semi-supervised learning task as follows.

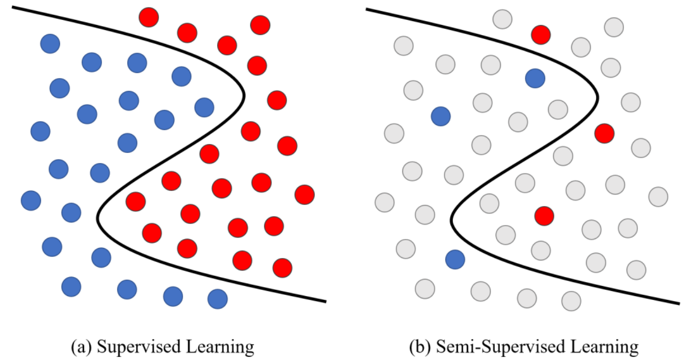


Fig. 3: Example comparison of supervised learning and semi-supervised learning.

Given a dataset $\mathcal{D}$ for training, we denote the labeled set with M labeled cases as $\mathcal{D}_L = \{x_i^l, y_i\}_{i=1}^M$, and the unlabeled set with N unlabeled cases as $\mathcal{D}_U = \{x_i^u\}_{i=1}^N$, where x_i^l and x_i^u denote the input images and y_i denotes the corresponding ground truth of labeled data. Generally, $\mathcal{D}_L$ is a relative small subset of the entire dataset $\mathcal{D}$, which means $M \ll N$. For semi-supervised segmentation settings, we aim at building a data-efficient deep learning model with the combination of $\mathcal{D}_L$ and $\mathcal{D}_U$ and making the performance to be comparable to an optimal model trained over fully labeled dataset.

Based on whether test data are wholly available in the training process, semi-supervised learning can be classified into two settings: the transductive learning and the inductive learning. For transductive learning, it is assumed that the unlabeled samples in the training process are exactly the data to be predicted (i.e. the test set), and the purpose of the transductive learning is to generalize the model over these unlabeled samples. While for inductive learning, the semi-supervised model will be applied to new unseen data.

2.2 Assumptions for Semi-Supervised Learning

For semi-supervised learning, an essential prerequisite is that the data distribution should be under some assumptions that the structure of the data remains constant. Otherwise, it is impossible to generalize from a finite training set to an infinite invisible set, where semi-supervised learning may not improve supervised learning and may even degrade the prediction accuracy by misleading inferences. The three basic assumptions for semi-supervised learning include:

The Smoothness Assumption. If two samples x_1 and x_2 are similar (e.g. in the same cluster), their corresponding outputs y_1 and y_2 should also be similar (e.g. belong to the same category), and vice versa.

The Cluster Assumption. This assumption refers to that if the sample in a single class tend to form a cluster, they belong to the same class cluster when the data points can be connected by short curves that do not pass through any low-density region. Therefore, the learning algorithm can use a large amount of unlabeled data to adjust the classification boundary.

The Manifold Assumption. If two samples x_1 and x_2 are located in a local neighborhood in the low-dimensional

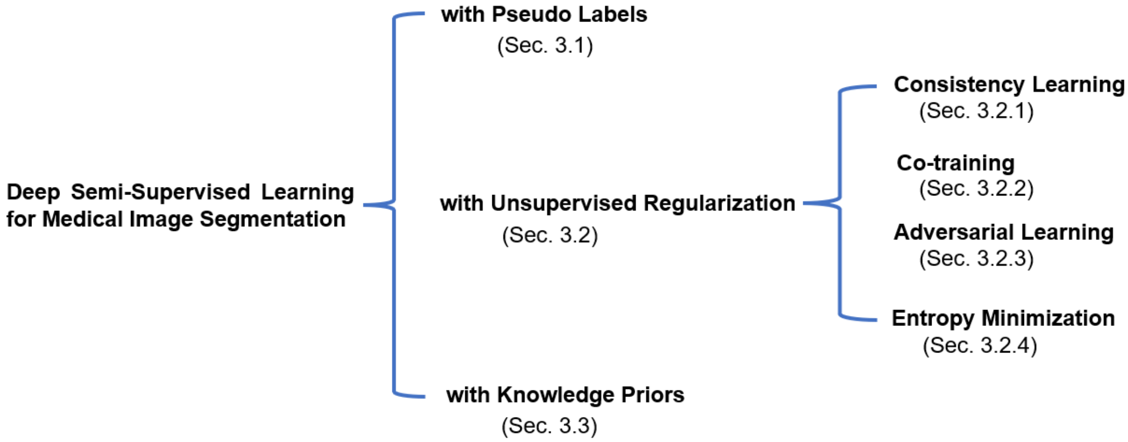


Fig. 4: The overview of existing deep semi-supervised learning methods for medical image segmentation.

manifold, they have similar class labels. This assumption reflects the local smoothness of the decision boundary.

2.3 Network Architectures for Medical Image Segmentation

The legendary U-Net [6], [7] has been widely used in various medical image segmentation tasks. The architecture of U-Net consists of an encoder with down-sampling layers and a decoder with up-sampling layers. The features of different scales are fused by concatenating the feature maps of the down-sampling layers and the corresponding up-sampling layers. Since the introduction of U-Net, many variants of encoder-decoder architecture have been proposed to improve it. Specifically, nnU-Net (no-new-U-Net) [13] has been proposed to automatically configure the pre-processing, the network architecture, the training, the inference, and the post-processing to a given dataset for medical image segmentation based on the encoder-decoder structure of U-Net. Without manual intervention, nnU-Net surpasses most existing approaches and achieves the state-of-the-art performance in several fully supervised medical image segmentation tasks.

3 Related Work on Semi-Supervised Medical Image Segmentation

In this section, we mainly divide these semi-supervised medical image segmentation methods into three strategies as follows:

1) semi-supervised learning with pseudo labels, where unlabeled images are firstly predicted and pseudo labeled by a segmentation model and then used as new examples for further training.

2) semi-supervised learning with unsupervised regularization, where unlabeled images are used jointly with labeled data to train a segmentation model with unsupervised regularization. This section mainly contains consistency learning, co-training, adversarial learning, entropy minimization.

3) semi-supervised learning with knowledge priors, where unlabeled images is utilized to enable the model with knowledge priors like the shape and position of the targets to improve the representation ability for medical image segmentation.

3.1 Semi-Supervised Medical Image Segmentation with Pseudo Labels

To utilize unlabeled data, a direct and intuitive method is assigning pseudo annotations for unlabeled images, and then using the pseudo labeled images in conjunction with labeled images to update the segmentation model. Pseudo labeling is commonly implemented in an iterative manner therefore the model can improve the quality of pseudo annotations iteratively. Algorithm 1 presents the overall workflow of this strategy.

Firstly, an initial segmentation model is trained using limited labeled data. The initial segmentation model is then applied to unlabeled data to generate pseudo segmentation masks. After that, labeled dataset is then merged with pseudo-labeled dataset to update the initial model. The training procedure alternates between the two steps introduced above, until a predefined iteration number.

Algorithm 1 Training procedure of semi-supervised learning with pseudo labels.

Input: $\{x^l, y^l\}$ from labeled dataset D_L, $\{x^u\}$ from unlabeled dataset D_U, initial segmentation model $\mathcal{M}_0$, iteration times $\mathcal{T}$
Output: Trained segmentation model $\mathcal{M}_\mathcal{T}$
Training initial segmentation model $\mathcal{M}_0$ with D_L
for $i \leftarrow 1$ to $\mathcal{T}$ **do**
  Generate pseudo labels $\{\hat{y}^u\}$ of unlabeled cases $\{x^u\}$ with model $\mathcal{M}_{i-1}$
  Generate new training dataset D_{PLi} with the combination of labeled dataset $\{x^l, y^l\}$ and pseudo labeled dataset with $\{x^u, \hat{y}^u\}$
  $\mathcal{M}_i \leftarrow$ Fine-tuning model $\mathcal{M}_{i-1}$ using D_{PLi}
end for
return Updated model $\mathcal{M}_\mathcal{T}$

Within this strategy for semi-supervised learning, these methods mainly differ in the model initialization, generation of pseudo labels, and how the noise in pseudo labels is handled. The outputs of an under-trained segmentation model with limited labeled data are noisy. If these noisy outputs are used as pseudo labels directly, it may make the training process unstable and hurt the performance [22]. For better leverage of the pseudo labels with potential noise, lots of methods have been proposed. In this section, we will explain the generation of pseudo-labels from two aspects: direct or indirect generation.

Pseudo labels from direct generation are mostly based on the predictions of a trained model in an online manner. A common method is to choose an unlabeled pixel with maximum predicted probability greater than the setting threshold. However, the predictions may be noisy and it is unreasonable to set the same threshold fit for all the samples. Based on the work in [22], [23], the pseudo labels with higher confidence are usually more effective. Therefore, many confidence- or uncertainty-aware methods are proposed to generate more stable and reliable pseudo labels. Yao *et al.* [24] propose a confidence-aware cross pseudo supervision network to improve the pseudo label quality of unlabeled images from unknown distributions. The KL-divergence of the predictions of the original and transformed images is calculated as the variance used for the proposed confidence-aware cross loss. Wang *et al.* [23] add a trust module to re-evaluate the pseudo labels from the model outputs and set a threshold to choose high confidence values. Except adding confidence-aware module, there are many other methods to improve the quality of pseudo labels. Li *et al.* [25] propose a self-ensembling strategy to build the up-to-date predictions via exponential moving average so as to avoid noisy and unstable pseudo-labels. Morphological methods and machine learning methods can be used to refine the pseudo labels. Superpixel maps calculated by simple linear iterative clustering (SLIC) algorithm [26] are introduced to refine pseudo labels in [27]. This algorithm is suitable for segmentation of targets with irregular shapes. Some algorithms add additional networks to further rectify the pseudo labels. Shi *et al.* [28] propose the conservative-radical network. The object conservative setting tends to predict pixels into background while the object radical setting tends to predict pixels into foreground. The certain region in predictions of unlabeled data is the overlap between conservative and radical settings and employed as pseudo labels. Zhang *et al.* [29] rectify the segmentation results of unlabeled data through another error segmentation network followed by the main segmentation network. The segmentation errors are divided into intra-class inconsistency or inter-class similarity problems. This method is appliable for different segmentation models and tasks.

Pseudo labels from indirect generation mostly are based on the label propagation *e.g.* prototype learning, nearest-neighbor matching [30], [31]. However, the indirect generation ways are time-consuming and demand higher memory comsumption, mostly in an offline manner. Han *et al.* [32] generate class representations from labeled data based on prototype learning. Through calculating the distances between feature vectors of unlabeled images and each class representation followed by a series morphological operations, high-quality pseudo labels are then generated.

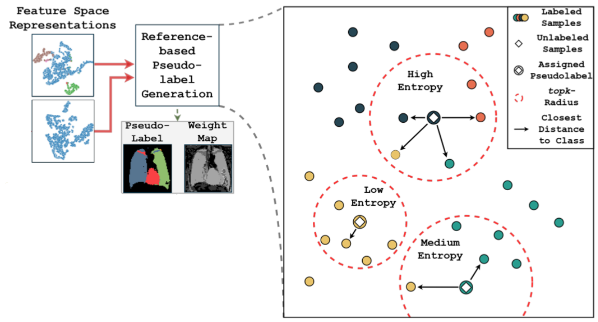


Fig. 5: Reference-guided Pseudo-Label Generation [31]

However, this prototype learning-based label propagation strategy requests high quality and representative feature extraction. Another common label propagation methods can premeditate the relations among data points of the whole dataset. Wang *et al.* [30] propose neighbor matching to generate pseudo-labels on a weight basis according to the embedding similarity with neighboring labeled data. [31] generate pseudo labels through transferring semantics that have a best fit with the unlabeled data in feature space among a pool of labeled reference images, as shown in Figure 5. In this way, confirmation bias which usually exists in network prediction-based pseudo label generation methods, can be avoided.

Along with adding more high-confidence pseudo labels, pseudo labeling encourages low-density separation between classes. The quality of pseudo labels is the main constraint for pseudo labeling strategy. The model is unable to correct its mistakes when it overfits to a small labeled data and has confirmation bias. The wrong predictions can be quickly amplified resulting in confident but erroneous pseudo labels with the training process [33]. Thus how to choose pseudo labels that will be added in the next training process and how many iterations to repeat need to be further considered.

3.2 Semi-Supervised Medical Image Segmentation with Unsupervised Regularization

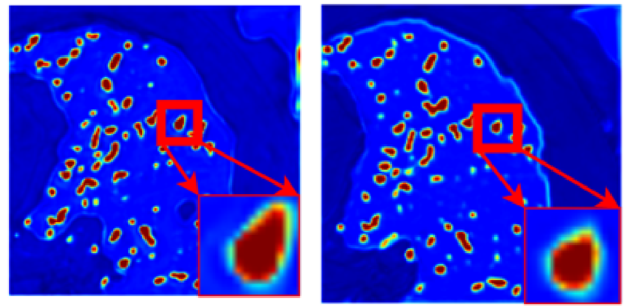
Different from generating pseudo labels and updating the segmentation model in an iterative manner, some recent progress in semi-supervised medical image segmentation has been focused on incorporating unlabeled data into the training procedure by generating a supervision signal with unsupervised regularization like unsupervised loss functions. Algorithm 2 presents the overall workflow of this strategy. Different choices of the unsupervised loss functions and regularization terms lead to different semi-supervised models. Generally, unsupervised regularization can be formulated into three sub-categories: consistency learning, co-training and entropy minimization.

3.2.1 Unsupervised Regularization with Consistency Learning

For unsupervised regularization, consistency learning is widely applied by enforcing an invariance of predictions of input images under different perturbations and pushing the decision boundary to low-density regions, based on the assumptions that the perturbations should not change the

Algorithm 2 Training procedure of semi-supervised learning with unsupervised regularization.

Input: $\{x^l, y^l\}$ from labeled dataset D_L, $\{x^u\}$ from unlabeled dataset D_U, segmentation model $\mathcal{M}$
Output: Trained segmentation model $\mathcal{M}$
while not converge **do**
    Calculate supervised segmentation loss $\mathcal{L}_{sup}(\theta; \mathcal{D}_L)$
    Calculate unsupervised loss $\mathcal{L}_{unsup}(\theta; \mathcal{D})$
    Update the segmentation model $\mathcal{M}$ with the combination of supervised loss $\mathcal{L}_{sup}$ and unsupervised loss $\mathcal{L}_{unsup}$
end while
return Trained segmentation model $\mathcal{M}$



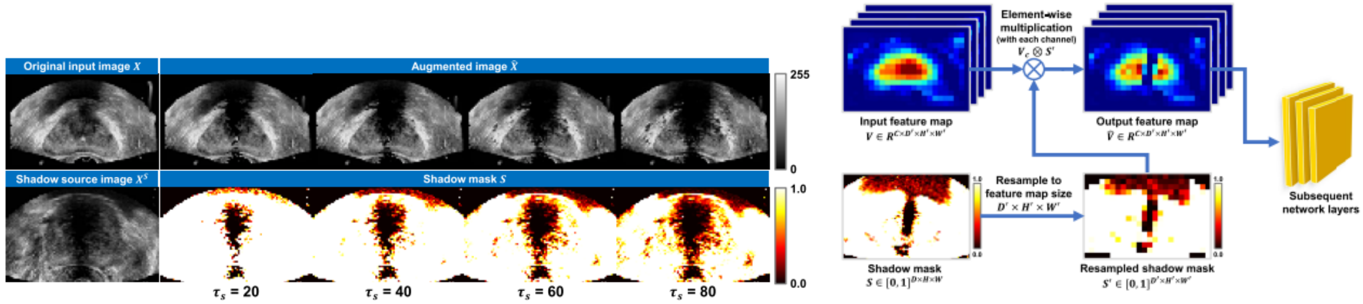
(a) Atrous convolution [42] to enlarge foreground features (b) Skip-connections [47] to shrink foreground features

Fig. 6: Morphological Feature Perturbations through Designing Different Network Architectures [41].

output of the model. The consistency between two objects can be calculated through Kullback-Leibler (KL) divergence, mean squared error (MSE), Jensen-Shannon divergence (JS) and so on. Such consistency learning-based methods are popular in semi-supervised medical image segmentation tasks due to their simplicity.

There are lots of perturbations, which can be divided into input perturbations and feature map perturbations. The perturbations should be meaningful for corresponding task and the effect of perturbations on segmentation performance has an upper bound, when adding more perturbations, the segmentation performance won't be further improved [34]. There are some commonly used input perturbations, such as Gaussian noise, Gaussian blurring, randomly rotation, scaling and contrast variations, and the segmentation network is encouraged to be transformation-consistent for unlabeled data [35]. Bortsova *et al.* [36] explore the equivariance to elastic deformations and encourage the segmentation consistency between the predictions of the two identical branches which receive differently transformed images. Huang *et al.* [34] add cutout content loss and slice misalignment as input perturbations. Another common consistency is mix-up consistency [37], [38], [39], which encourages the segmentation of interpolation of two data to be consistent with the interpolation of segmentation results of those data. Apart from disturbances on inputs, there are also many studies focusing on disturbances at feature map level. Zheng *et al.* [40] propose to add random noise to the parameter calculations of the teacher model. Xu *et al.* [41] propose morphological feature perturbations through designing different network architectures, as shown in Figure 6, Atrous convolutions can enlarge foreground features while skip-connections will shrink foreground features [42], [43]. Li *et al.* [44] add seven types of feature perturbations to seven extra decoders and require this seven predictions to be consistent with the main decoder. These feature level perturbations are feature noise, feature dropout, object masking, context masking, guided cutout, intermediate VAT, and random dropout, based on the work in [45]. There are also studies that applying perturbations both at the input and feature map levels. Xu *et al.* [46] propose a novel shadow consistency which contains shadow augmentation 7(a) and shadow dropout 7(b) to simulate the low image quality and shadow artifacts in medical images. Specifically, shadow augmentation is a perturbation through adding simulated shadow artifacts to the input images while shadow dropout will drop neural nodes according to the prior knowledge of the shadow artifacts, which is a disturbance acting directly on feature maps. Note that if the perturbations are too weak, it may cause the Lazy Student Phenomenon, but large perturbations may confuse the teacher and student and lead to low performance. Shu *et al.* [38] add a transductive monitor for further knowledge distillation to narrow the semantic gap between the student model and teacher model.

Instead of adding perturbations, there are also some different consistency learning methods. For instance, Sajjadi *et al.* [48] propose Π Model to create two random augmentations of a sample for both labeled and unlabeled data. In the training process, the model expects the output of same unlabeled sample propagates forward twice under different random perturbations to be consistent. Samuli *et al.* [49] propose temporal ensembling strategy to use exponential moving average (EMA) predictions for unlabeled data as the consistency targets. However, maintaining the EMA predictions during the training process is a heavy burden. To issue the problem, Tarvainen *et al.* [50] propose to use a teacher model with the EMA weights of the student model for training and enforce the consistency of predictions from perturbed inputs between student and teacher models. Zeng *et al.* [51] improve the EMA weighted way in teacher models. They add a feedback signal form the performance of the student on the labeled set, through which the teacher model can be updated by gradient descent algorithm autonomously and purposefully. However, due to the limited number of labeled data, the predictions of the teacher model can be wrong at some locations and might confuse the student model. So the uncertainty or confidence estimation are utilized to learn from more meaningful and reliable targets during training. Yu *et al.* [52] extend the mean teacher paradigm with an uncertainty estimation strategy through Monte Carlo dropout [53]. Xie *et al.* [54] add a confidence-aware module to learn the model confidence under the guidance of labeled data. Luo *et al.* [55], [56] calculate uncertainty using pyramid predictions in one forward pass and proposed an multi-level uncertainty rectified pyramid consistency regularization. Fang *et al.* [57] attach an error estimation network to predict the CE loss map of the teacher's prediction. Then the consistency loss will be calculated on



(a) Shadow Augmentation

(b) Shadow Dropout

Fig. 7: Shadow Augmentation and Dropout [46]

low CE loss pixels. Zhao *et al.* [58] introduce cross-level consistency constraint which is calculated between patches and the full image. Except encouraging consistency on network segmentation results directly, generative consistency [59] is proposed through a generation network that reconstructs medical images from its predictions of the segmentation network. Xu *et al.* [60] propose contour consistency and utilize Fourier series which contained a series of harmonics as an elliptical descriptor. Through minimizing the L2 distance of the parameters between the student and the teacher branch, the model is equipped with shape awareness. However, this method needs to choose different maximum harmonic numbers for the segmentation of targets with different irregularity. Chen *et al.* [61] propose multi-level consistency loss which computes the similarities between multi-scale features in an additional discriminator, where the inputs are the segmentation regions by multiplying the unlabeled input image with predicted segmentation probability maps instead of segmentation probability maps. Hu *et al.* [62] propose attention guided consistency which encourages the attention maps from the student model and the teacher model to be consistent. Each image contains the same class object, so different images share similar semantics in the feature space. Xie *et al.* [63] introduce intra- and inter-pair consistency to augment feature maps. The pixel-level relation between a pair of images in the feature space is first calculated to obtain the attention maps that highlight the regions with the same semantics but on different images. Then multiple attention maps are taken into account to filter the low-confidence regions and then merged with the original feature map to improve its representation ability. Liu *et al.* [64] propose contrastive consistency which encourages segmentation outputs to be consistent in class-level through foreground and background class-vectors generated from a classification network. Xu *et al.* [65] propose the cyclic prototype consistency learning (CPCL) framework which contains a labeled-to-unlabeled (L2U) prototypical forward process and an unlabeled-to-labeled (U2L) backward process. The L2U forward consistency can transfer the real label supervision signals to unlabeled data while the U2L backward consistency can directly use the labeled data to guide the learning from unlabeled data, thus turning ”unsupervised” consistency into “supervised” consistency.

Other than utilizing data-level perturbations for consistency learning, some methods focus on building task-

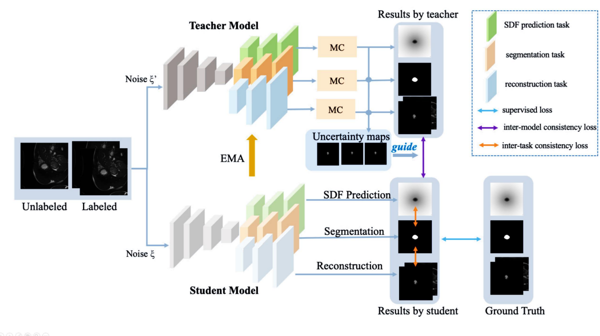


Fig. 8: Tripled-uncertainty Guided Mean Teacher Model [72]

level regularization by adding auxiliary task to leverage geometric information. Li *et al.* [66] develop a multi-task network to build shape-aware constraints with adversarial regularization. Liu *et al.* [67] propose a shape-aware multi-task framework which contained segmentation, Signed Distance Map prediction and Organ Contour prediction. Luo *et al.* [68] combine the level set function regression task with the segmentation task to form a dual-task consistency for semi-supervised learning. Zhang *et al.* [69] propose dual-task mutual learning framework by encouraging dual-task networks to explore useful knowledge from each other. Based on dual-task framework, Zhang *et al.* [70] utilize both segmentation task and regression task for self-ensembling and utilize estimated uncertainty to guide the mutual consistency learning and obtain further performance improvement. Chen *et al.* [71] propose a dual-task consistency joint learning framework that encouraged the segmentation results to be consistent with the transformation of the signed distance map predictions. Wang *et al.* [72] inject multi-task learning into mean teacher architecture which contain the segmentation task, the reconstruction task, and the SDF prediction task so that the model can take account of the data-, model- and task-level consistency, as shown in Figure 8. Besides, they propose an uncertainty weighted integration (UWI) strategy to estimate the uncertainty on all tasks and develop a triple-uncertainty based on these tasks to guide the student model to learn reliable information from teacher.

3.2.2 Unsupervised Regularization with Co-Training

Based on the work in [73], co-training framework assumes that each data has two different views and each view has sufficient information that can give predictions independently. One view is redundant to another view and the model is encouraged to have consistent predictions on these two views. It first learns a separate segmentation model for each view on labeled data, and then the predictions of the two models on unlabeled data are gradually added to training set to continue the training. Note that different from self-training methods, co-training methods add pseudo labels from one view to the training set in other views. And the difference between co-training and consistency learning is that the pseudo labels of unlabeled data will act as supervised signals to train other models while the consistency learning encourages the outputs for different perturbations to be consistent. That is to say, all of the models in co-training will be updated through gradient descent algorithm while there is only one main model in consistency learning updated through gradient descent algorithm.

The core of co-training is how to construct two (or more) deep models of approximately represent sufficiently independent views. The methods mainly contain using different sources of data, employing different network architectures and using special training methods to obtain diverse deep models. For medical images, the data can be from different modalities or medical centers leading to different distributions. Zhu *et al.* [74] propose a co-training framework for unpaired multi-modal learning. This framework contains two segmentation networks and two image translation networks across two modalities. They utilize the pseudo-labels (from unlabeled data) or labels (from labeled data) from one modality to train the segmentation network in the other modality after image translation. For one thing, it increases supervised signals. For another, it adds modality-level consistency. Chen *et al.* [75] leveraged unpaired multi-modality images to be cross-modal consistent in anatomy and semantic information. The multi modalities which are collaborative and complementary could encourage better modality-independent representation learning. Liu *et al.* [76] present a co-training framework for domain-adaptive medical image segmentation. This framework contains two segmentors used for semi-supervised segmentation task (labeled and unlabeled target domain data as inputs) and unsupervised domain adaptation task (labeled source domain data and unlabeled target domain data as inputs), respectively. As different models usually extract different representations, the different models in co-training framework can focus on different views. Except using CNN as the backbones, there are also some transformer-based backbones [77], [78]. As shown in Figure 9, Luo *et al.* [79] introduce the cross teaching between CNN and Transformer which implicitly encourages the consistency and complementary between different networks. Liu *et al.* [77] combine CNN blocks and Swin Transformer blocks as the backbone. Xiao *et al.* [78] add another teacher model with the transformer-based architecture. The teacher models communicate with each other with consistency regularization and guide the student learning process. However, when there are only one source of data available, training two (or more) identical networks may

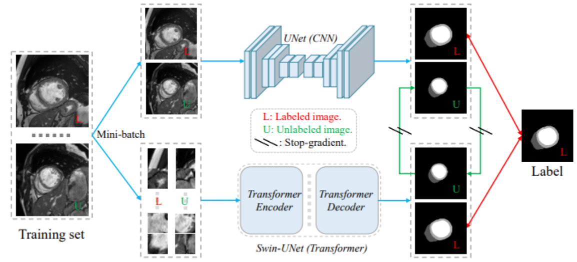


Fig. 9: Cross Teaching between CNN and Transformer [79]

lead to collapsed neural networks as the predictions from these models are encouraged to be similar. [80], [81] generate adversarial examples as another view. [82], [83] use coronal, sagittal and axial views of 3D medical images as view difference at input level and [82] also use asymmetric 3D kernels with 2D initialization as view difference at feature level. Except that, diverse deep models can also be trained using special training methods. For instance, Chen *et al.* [84] use output smearing to generate different labeled data sets to initialize diverse models. To maintain the diversity in the subsequent training process, the modules are fine-tuned using the generated sets in specific rounds.

Except constructing diverse models of sufficiently independent views, another line of researches focus on which pseudo label to choose in the subsequent training process. Although consistent predictions are encouraged across the networks, they may contain noise leading to unstable training process. An Uncertainty-Aware co-training framework [82] is proposed through estimating the confidence of each view via Bayesian uncertainty estimation. Wang *et al.* [85] develop a self-paced and self-consistent co-training framework. The self-paced strategy can encourage the network to transfer the knowledge of easier-to-segment regions to the harder ones gradually through minimizing a generalized Jensen-Shannon Divergence. Another way to alleviate the influence from noisy pseudo labels is through exponential mix-up decay to adjust the contribution of the supervision signals from both labels and pseudo labels across the training process [76].

3.2.3 Unsupervised Regularization with Adversarial Learning

Other than consistency learning, some researches use adversarial methods to encourage the segmentation of unlabeled images to be closer to those of the labeled images. These methods always contain a discriminator to distinguish the inputs from labeled annotations or unlabeled predictions [61], [93], [94], [95]. Zhang *et al.* [93] introduce adversarial learning to encourage the segmentation output of unlabeled data to be similar with the annotations of labeled data. Chen *et al.* [61] add a discriminator following the segmentation network which is used to distinguish between the input signed distance maps from labeled images or unlabeled images. Peiris *et al.* [95] add a critic network into the segmentation architecture which can perform the min-max game through discriminating between prediction masks and the ground truth masks. The experiments show that it could sharpen boundaries in prediction masks. The discriminator

TABLE 1: Representative works and empirical results on semi-supervised LA MRI segmentation benchmark.

Method	Highlights	$\mathcal{D}_L/\mathcal{D}_U$	Dice	Publication&Year
Baseline V-Net [86]	Fully supervised baseline with only labeled data	8/0	79.99	——
		16/0	86.03	
Upper-bound V-Net [86]	Fully supervised upper bound with all annotations	80/0	91.14	
UA-MT [52]	Teacher-student framework with the guidance of uncertainty	8/72	84.25	MICCAI 2019
		16/64	88.88	
SASS [66]	Incorporating signed distance maps for shape regularization	8/72	87.32	MICCAI 2020
		16/64	89.54	
DUWM [87]	Utilizing both segmentation and feature uncertainty	8/72	85.91	MICCAI 2020
		16/64	89.65	
LG-ER-MT [88]	Entropy minimization to produce high-confident predictions and local structural consistency to encourage inter-voxel similarities	8/72	85.54	MICCAI 2020
		16/64	89.62	
DTC [68]	Encourage the consistency between output segmentation maps and signed distance map	16/64	89.42	AAAI 2021
PDC-Net [89]	Parameter decoupling to encourage consistent predictions from two branch network	8/72	86.55	ICMV 2021
		16/64	89.76	
HCR-MT [90]	Teacher-student framework with multi-scale deep supervision and hierarchical consistency regularization	16/64	90.04	EMBC 2021
DTML [69]	Mutual learning of dual-task networks for generating segmentation and signed distance maps	16/64	90.12	PRCV 2021
MC-Net [91]	Consistency learning between outputs from two different decoders	8/72	87.71	MICCAI 2021
		16/64	90.34	
CASS [64]	Contrastive consistency on class-level	8/72	86.51	CMIG 2022
		16/64	89.81	
SimCVD [92]	Contrastive distillation of voxel-wise representation with signed distance maps	8/72	89.03	TMI 2022
		16/64	90.85	
CMM [38]	Asynchronously perform Cross-Mix Teaching and Transductive Monitor for active knowledge distillation	8/72	85.92	TMM 2022
		16/64	90.03	
DTCJL [71]	Semi-supervised dual-task consistent joint learning framework with task-level regularization	16/64	90.32	TCBB 2022

can also be used to generate pixel-wise confidence maps and select the trustworthy pixel predictions used for consistency learning. Wu *et al.* [96] add two discriminators for predicting confidence maps and distinguishing the segmentation results from labeled or unlabeled data. Through adding another auxiliary discriminator, the under trained primary discriminator due to limited labeled images can be alleviated. Li *et al.* [97] employ the U-net as the encoder and a conditional GAN as the decoder. Through reconstructing images from the predictive result of the encoder, the encoder can estimate the distribution of segmentation maps. Nie *et al.* [98] propose to adversarially train the segmentation network based on the confidence map from the confidence network and a region-attention based semi-supervised learning strategy to utilize unlabeled data for training. Hou *et al.* [99] add a leaking GAN into the semi-supervised framework which can pollute the discriminator by leaking information from the generator for more moderate generations. Chaitanya *et al.* [100] propose a novel task-driven data augmentation method to synthesize new training examples, where a generative network explicitly applies deformation fields and additional strength masks to model shape and strength changes. However, adversarial training may be challenging in terms of convergence.

3.2.4 Unsupervised Regularization with Entropy Minimization

Based on the assumption in semi-supervised learning that the decision boundary should lie in low-density regions, entropy minimization encourages the model to output low-entropy predictions on unlabeled data and avoids the class overlap. Therefore, semi-supervised learning algorithms are usually combined with entropy minimization [101], [102], [103]. Base on the work in [102], a loss term is added to minimize the entropy of the predictions of the model on unlabeled data and the object function turns to be 1.

$$\begin{aligned} C(\theta,\lambda;\mathcal{L}_n) &= L(\theta;\mathcal{L}_n) - \lambda H_{emp}(Y|X,Z;\mathcal{L}_n) \\ &= \sum_{i=1}^{n}\log(\sum_{k=1}^{K} z_{ik} f_k(X_i)) \\ &+ \lambda \sum_{i=1}^{n}\sum_{k=1}^{K} g_k(x_i,z_i)\log g_k(x_i,z_i) \end{aligned} \tag{1}$$

where $L(\theta;\mathcal{L}_n)$ is the conditional log-likelihood and sensitive to the labeled data, and $H_{emp}(Y|X,Z;\mathcal{L}_n)$ is conditional entropy and only affected by the unlabeled data which works to minimize the class overlap. Wu *et al.* [101] add entropy minimization technique in the student branch.

Berthelot *et al.* [37] propose MixMatch to use a sharpening function on the target distribution of unlabeled data to minimize the entropy. The sharpening through adjusting the "temperature" of this categorical distribution is as follow:

$$Sharpen(p,T)_i = p_i^{\frac{1}{T}} \Big/ \sum_{j=1}^{L} p_j^{\frac{1}{T}} \tag{2}$$

where T is a hyperparameter. As $T \rightarrow 0$, the output of $Sharpen(p,T)$ will approach a Dirac ("one-hot") distribution. Lowering temperature encourages model to produce lower-entropy predictions. However, the hyperparameter needs to be set carefully and different samples may have different T, so [91] propose an adaptive sharpening which can adjust T adaptively for each sample according to its uncertainty predicted by the model. [104] introduce a mutual exclusivity loss for multi-class problems that explicitly forces the predictions to be mutually exclusive and encourages the decision boundary to lie on the low density space between the manifolds corresponding to different classes of data, which has a better performance in object detection task compared with entropy minimization in [102].

Another application of Entropy Minimization is the use of hard label in the pseudo labeling. As arg max operation applied to a probability distribution can produce a valid "one-hot" low-entropy (i.e., high-confidence) distribution, both the Entropy Minimization and pseudo labeling encourages the decision boundary passing low-density regions. Therefore, the strategy of using hard label in the pseudo labeling is closely related with Entropy Minimization [105]. However, a high capacity model that tends to overfit quickly can give high-confidence predictions which also have low entropy [106]. Therefore, Entropy Minimization doesn't work in some cases [102]. However, when combined with other semi-supervised learning strategies, Entropy Minimization may improve the performance [107].

3.3 Semi-Supervised Medical Image Segmentation with Knowledge Priors

Knowledge priors are the information that a learner already has before it learns new information, and sometimes are helpful for dealing with new tasks. Compared with non-medical images, medical images have many anatomical priors such as the shape and position of organs and incorporating the anatomical prior knowledge in deep learning can improve the performance for medical image segmentation [108]. Some semi-supervised algorithms utilize knowledge priors to improve the representation ability for the new task.

Self-supervised pretraining is an application of prior knowledge. As there are large unlabeled data in semi-supervised learning, the model can learn useful representations and visual priors through an efficient proxy task pretraining. Huang *et al.* [34] add a reconstruction pre-training from the counterparts to avoid the network being randomly initialized in a cold start stage. Huang He *et al.* [109] pretrain an auto-encoder through a reconstruction proxy task and the deep prior anatomy (DPA) features extracted from it are then embedded for segmenting thin structures and large inter-anatomy variation, as shown in Figure 10. Hu *et al.* [110] introduce the self-supervised image-level and

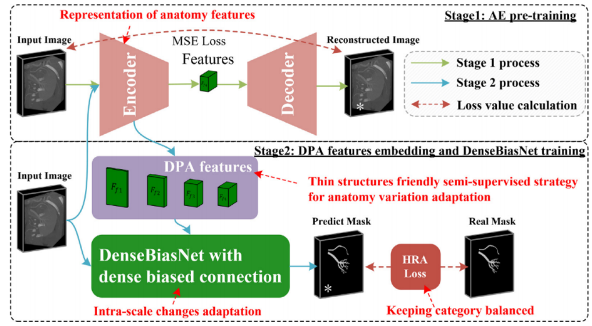


Fig. 10: DPA-DenseBiasNet for Fine Renal Artery Segmentation [109]

supervised pixel-level contrastive pre-training into the semi-supervised framework. Wang *et al.* [111] use superpixel to separate an image into regions and learned intra- and inter-organ representation based on contrastive learning, then the model is used to initialize the semi-supervised framework, which boost the performance significantly. Except self-supervised pretraining, proxy tasks and contrastive loss can also be utilized into semi-supervised training process as regularization. Yang *et al.* [112] introduce self-supervised jigsaw puzzle task into the semi-supervised training process for better feature representation. [113] propose a dual-task network with a shared encoder and two independent decoders for segmentation and lesion region inpainting. Peng *et al.* [114] integrate semi-supervised learning with self-paced contrastive learning, which can assign an importance weight to the specific loss of each positive pair based on meta-labels, different from unsupervised contrastive loss. Wu *et al.* [101] add patch- and pixel-level dense contrastive loss to align the features from the teacher and student models. They also add entropy minimization technique in the student branch. Zhao *et al.* [83] introduce the multi-scale multi-view global-local contrastive learning into co-training framework.

$$Confidence = \exp\left(-\frac{(PA - s_{mask})^2 + (s_{output} - s_{mask})^2}{2\sigma^2}\right) \tag{3}$$

$$s_{mask} = [s_{output} + 0.5] \tag{4}$$

Except from self-supervised learning, the following are other applications of prior knowledge for semi-supervised learning. The atlas map, as shown in Figure 11, which indicates the probability of the object appearing at some location, is widely applied in medical image segmentation [108], [115], [116], [117]. The targets need to be registered to a referenced volume. Then the probabilistic atlas (PA) can be generated through averaging manually masks after deformable of all annotated volumes. Zheng *et al.* [108] calculate the organ PA through averaging the manually segmented liver masks after registration of all annotated volumes and predefined the hard samples with the atlas values close to 0.5. Huang *et al.* [117] utilize PA to give the unlabeled data segmentation pixel-wise confidence (3 and 4) to select reliable pixel results. As can be seen in Figure 12, the confidence decreases from red to blue and the confidence is higher, when both PA and s_{output} are close to s_{mask},

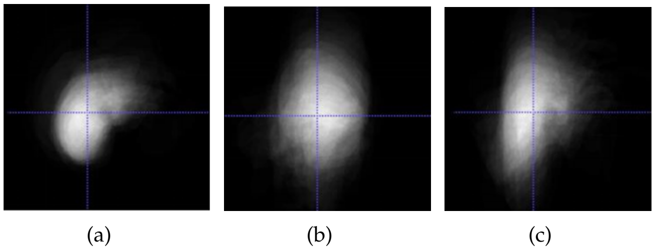


Fig. 11: The 3D Probabilistic atlas of liver organ [117], (a)-(c) are superior–inferior, left–right direction and anterior–posterior direction correspondingly.

that is 0 or 1. The method takes prior shape and position information into account, but it may be unsuitable for the target that has large positional variance. Li *et al.* [118] integrate contextual refinement into deformable registration-based segmentation processes in a semi-supervised learning paradigm, which only utilize an atlas image and a small amount of labeled data.

3.4 Other Semi-Supervised Medical Image Segmentation Methods

A frequently encountered obstacle in medical imaging is that, in real-world applications, the acquired data and annotations may be difficult to meet the assumptions, thus affecting the performance of semi-supervised learning. Other than these methodological developments for semi-supervised segmentation methods mentioned above, We have also compiled some different concerns in real-world applications.

As there is usually a large amount of unlabeled data in semi-supervised learning, the distribution of labeled and unlabeled data may be misaligned. For better leverage of large scale data from different distributions or medical centers, some methods are proposed to deal with distribution misalignment. Zhang *et al.* [119] try to align labeled data distribution and unlabeled data distribution through minimising the L2 distance between the feature maps of them. Meanwhile, to remain discriminative for the segmentation of labeled and unlabeled data, further segmentation supervision is obtained through comparing the non-local semantic relation matrix in feature maps from the ground truth label mask and the student inputs. Another work in [120] propose adaptive hierarchical dual consistency to use the dataset from different centers, which learns mapping networks adversarially to align the distributions and extend consistency learning into intra- and inter-consistency in cross-domain segmentation. Another idea for using data form multi centers is through meta-learning. Based on the work in [121], one distinct task is formulated for each medical centre such that a segmentation task is performed for a centre with labelled data while the contrastive learning task is performed for one with unlabelled data.

Another concern in semi-supervised learning is how to fuse different supervision signals for label-efficient semi-supervised learning. As existing public imaging datasets usually have different annotations for different tasks, like CT images singly labelled tumors or partially labelled organs. Zhang *et al.* [122] propose a dual-path semi-supervised conditional nnU-Net that can be trained on a union of partially labelled datasets, segmentation of organs at risk or tumors. Another situation is the integration of different levels of supervision signals. [123] propose multi-label deep supervision in semi-supervised framework, which leveraged image-level, box-level and pixel-level annotations. If only image-level or box-level labels exist, the pseudo labels would be constrained to the classes contained in that or to lie within coarse regions. Except that, the noisy pseudo labels generated from the teacher model is smoothed using max-pooling to match different level predictions from the decoder for multi-level consistency.

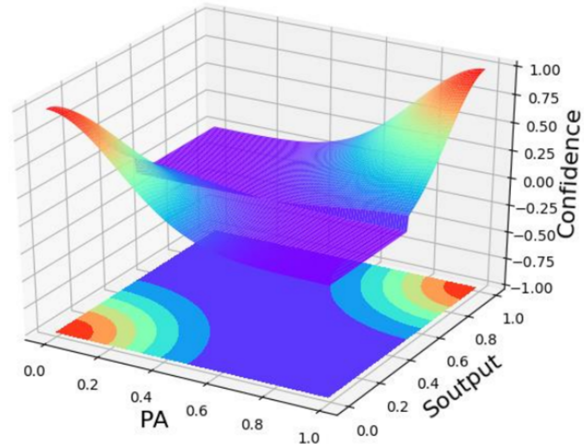


Fig. 12: Illustration of confidence map and its corresponding contour map with $\sigma = 0.25$ [117]

Another common problem in segmentation is class imbalance. In semi-supervised learning, class imbalance and limited labeled data may further bring the confirmation bias and uncertainty imbalance problem. Lin *et al.* [124] propose a dual uncertainty-aware sampling strategy to sample low-confident categories of pixels for unsupervised consistency learning. Wang *et al.* [23] add a re-weighting module calculated by the pixel proportion of classes in the labeled training set and pseudo-label training set to the cross entropy loss for dealing with class imbalance.

Besides, most of previous semi-supervised frameworks are discriminative models, where labeled data is only used in the early training stage and the model may tend to overfit to the labeled data [125]. Wang *et al.* [125] proposed a Bayesian deep learning framework for semi-supervised segmentation. In that way, both labeled and unlabeled data are utilized to estimate the joint distribution, which alleviates potential overfitting problem caused by using labeled data for early training only.

4 Analysis of Empirical Results for Semi-Supervised Medical Image Segmentation

4.1 Common Evaluation Metrics for Medical Image Segmentation

For medical image segmentation tasks, Dice Similarity Coefficient (DSC) is a widely used metric to measure the region overlap ratio of the ground truth G and segmentation result

S. Another similar metric IoU (or Jaccard) is used as an alternative for the evaluation. These two metrics are defined as follows:

$$DSC = \frac{2|G \cap S|}{|G| + |S|}, \quad IoU = \frac{|G \cap S|}{|G \cup S|}. \tag{5}$$

However, region-based metrics like DSC cannot well reflect the boundary error or small region of mis-segmentation. To issue this limitation, boundary-based evaluation metrics like Hausdorff Distance (HD) are applied to focus on the boundary distance error defined as follows:

$$HD(\partial G, \partial S) = \max(\max_{x \in \partial G} \min_{y \in \partial S} ||x-y||_2, \max_{x \in \partial S} \min_{y \in \partial G} ||x-y||_2), \tag{6}$$

where ∂G and ∂S represent the boundary of the ground truth and the segmentation result, respectively. To eliminate the influence caused by small subsets of outliers, 95% Hausdorff Distance (95HD) is also widely used, which is based on the calculation of the 95th percentile of the distances between boundary points.

4.2 Benchmark Datasets for Semi-Supervised Medical Image Segmentation

In addition to the promising progress in semi-supervised medical image segmentation methods, several segmentation benchmarks are also evolved to ensure a fair comparison of these methods with the same task setting on same public dataset.

LA dataset. The LA benchmark dataset [126] from the Left Atrium Segmentation Challenge [2] contains 100 3D gadolinium-enhanced MR imaging scans (GE-MRIs) for training, with an isotropic resolution of $0.625 \times 0.625 \times 0.625mm^3$. Since the testing set of LA does not include public annotations, for the settings in [52], the 100 training scans are splitted into 80 scans for training and 20 scans for testing. Out of the 80 training scans, 20% (i.e. 16 scans) are used as labeled data and the remaining as unlabeled data. V-Net [86] is used as the network backbone for all experiments with a joint cross-entropy loss and dice loss for training. For supervised comparisons, V-Net trained with only labeled data (i.e. 16 scans) and trained with all labeled data (i.e. 80 scans) is performed as lower bound and upper bound for semi-supervised learning. As shown in Table. 1, as one of the most popular benchmark dataset for semi-supervised medical image segmentation, many methods are further proposed and evaluated on the same dataset under the same task settings following the task design of [52]. Specifically, several researches further promote the benchmark with 10% (i.e. 8 scans) labeled scans to further evaluate the performance under the circumstance with fewer labeled data.

Pancreas CT dataset. The NIH Pancreas CT segmentation dataset [127] contains 82 3D abdominal contrast-enhanced CT volumes, which are collected from 53 male and 27 female subjects at the National Institutes of Health Clinical Center [3]. The dataset are collected on Philips and Siemens MDCT scanners and have a fixed resolution of 512×512 with varying thicknesses from 1.5 to 2.5 mm, while the axial view slice number can vary from 181 to 466. In [82], the dataset is randomly split into 20 testing cases and 62 training cases. Experimental results with 10% labeled training cases (6 labeled and 56 unlabeled) and 20% labeled training cases (12 labeled and 50 unlabeled) is reported. Following the pre-processing in [128], the voxel values are clipped to the range of [-125,275] Hounsfield Units (HU) and further re-sampled to an isotropic resolution of $1 \times 1 \times 1mm^3$.

BraTS dataset. The Brain Tumor Segmentation (BraTS) 2019 dataset [129] contains multi-institutional preoperative MRI of 335 glioma patients, where each patient has four modalities of MRI scans including T1, T1Gd, T2 and T2-FLAIR with neuroradiologist-examined labels. For several existing approaches [55], [65], [70], T2-FLAIR for whole tumor segmentation is used since such modality can better manifest the malignant tumors [130]. All the scans are re-sampled to the same resolution of $1 \times 1 \times 1mm^3$ with intensity normalized to zero mean and unit variance. For semi-supervised settings, the dataset is splitted into 250 scans for training, 25 scans for validation and the remaining 60 scans for testing. Among the 250 training scans, two different settings are performed with 10%/25 and 20%/50 scans as labeled data and the remaining scans as unlabeled data.

ACDC dataset. The ACDC (Automated Cardiac Diagnosis Challenge) dataset [131] was collected from real clinical exams acquired at the University Hospital of Dijon [4]. The dataset contains multi-slice 2D cine cardiac MR imaging samples from 100 patients for training. For semi-supervised settings, the dataset is splitted into 70 scans for training, 10 scans for validation and 20 scans for testing. Unlike previous 3D binary segmentation benchmark datasets, ACDC is a 2D multi-class segmentation task including RV cavity, myocardium and the LV cavity.

5 Existing Challenges and Future Directions

Although considerable performance has been achieved for semi-supervised medical image segmentation tasks, there are still several open questions for future work. In this section, we outline some of these challenges and future directions as follows.

Misaligned distribution and class imbalance. As described in Section 4.2, existing semi-supervised medical image segmentation approaches have achieved comparable results with upper-bound fully supervised results in some benchmark datasets like LA segmentation [126]. However, these benchmarks are relatively "simple" tasks, with small amount of experimental data where the training and test set are from the same domain/medical center. However, a clinical applicable deep learning model should be generalized suitably across multiple centres and scanner vendors from different domains [132]. As there is usually a large amount of unlabeled data in semi-supervised learning, the distribution of labeled and unlabeled data may be misaligned. This limitation is also highlighted by recent semi-supervised medical segmentation benchmarks like [133] and

2. http://atriaseg2018.cardiacatlas.org/data/
3. https://wiki.cancerimagingarchive.net/display/Public/Pancreas-CT
4. https://www.creatis.insa-lyon.fr/Challenge/acdc/databases.html

FLARE 22 challenge [5]. Based on the work in [106], adding unlabeled data from a mismatched distribution from labeled data can lower the performance compared to not using any unlabeled data. Therefore, it is of great importance to issue the challenge of misaligned distribution for semi-supervised learning. As for class imbalance, when the training data is highly imbalanced, the trained model will show bias towards the majority classes, and may completely ignore the minority classes in some extreme cases [134]. Besides, for semi-supervised multi-class segmentation, there usually exists the uncertainty imbalance problem brought by class imbalance and limited labeled data. Recent studies [135] found that aleatoric uncertainty derived from the entropy of the predictions may lead to sub-optimal results in a multi-class context.

Methodological analysis. Existing semi-supervised medical image segmentation approaches predominantly use unlabeled data to generate constraints, then the models are updated with supervised loss for labeled data and unsupervised loss/constraints for unlabeled data (or both labeled and unlabeled data). Generally, there is only a single weight to balance between supervised and unsupervised loss as described in many approaches [52], [68], [69]. In other words, all the unlabeled data are treated equally for semi-supervised learning. However, not all unlabeled data is equally appropriate for the learning procedure of the model. For example, when the estimation of an unlabeled case is incorrect, training on that particular label-estimate may hurt the overall performance. To issue this problem, it is important to encourage the model focusing on more challenging areas/cases and therefore exploit more useful information from unlabeled data like assigning different weights for each unlabeled example [136]. Recent studies [137] also found that the quality of the perturbations is key to obtaining reasonable performances for semi-supervised learning, especially in the case of efficient data augmentations or perturbations schemes when the data lies in the neighborhood of low-dimensional manifolds.

Integration with other annotation-efficient approaches. For existing semi-supervised learning approaches, we still need a small amount of well-annotated labeled data to guide the learning of unlabeled data. However, acquiring such fully annotated training data can still be costly, especially for the tasks of medical image segmentation. To further alleviate the annotation cost, some researches integrate semi-supervised learning with other annotation-efficient approaches like utilizing partially labelled datasets [122], leveraging image-level, box-level and pixel-level annotations [123] or scribble supervisions [138], or exploiting noisy labeled data [139].

6 Conclusion

Semi-supervised learning has been widely applied to medical image segmentation tasks since it alleviates the heavy burden of acquiring expert-examined annotations and takes the advantage of unlabeled data which is much easier to acquire. In this survey, we provide a taxonomy of existing deep semi-supervised learning methods for medical image segmentation tasks and group these methods into three main categories, namely, pseudo labels, unsupervised regularization, and knowledge priors. Other than summarizing technical novelties of these approaches, we also analyse and discuss the empirical results of these methods on several public benchmark datasets. Furthermore, we analysed and discussed the limitations and several unsolved problems of existing approaches. We hope this review could inspire the research community to explore solutions for this challenge and further promote the developments in this impactful area of research.

Acknowledgments

This paper was supported in part by the National Science Foundation under Grant 32000687, and in part by the University Synergy Innovation Program of Anhui Province under Grant GXXT-2019-044 and Beijing Natural Science Foundation under Grant Z200024. We also appreciate the efforts of literature collection and code implementations of SSL4MIS [6] and several public benchmarks.

References

[1] B. Van Ginneken, C. M. Schaefer-Prokop, and M. Prokop, "Computer-aided diagnosis: how to move from the laboratory to the clinic," *Radiology*, vol. 261, no. 3, pp. 719–732, 2011.

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

[3] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *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*, vol. 67, p. 101821, 2021.

[4] V. Oreiller, V. Andrearczyk, M. Jreige, S. Boughdad, H. Elhalawani, J. Castelli, M. Vallières, S. Zhu, J. Xie, Y. Peng *et al.*, "Head and neck tumor segmentation in pet/ct: the hecktor challenge," *Medical image analysis*, vol. 77, p. 102336, 2022.

[5] A. Lalande, Z. Chen, T. Pommier, T. Decourselle, A. Qayyum, M. Salomon, D. Ginhac, Y. Skandarani, A. Boucher, K. Brahim *et al.*, "Deep learning methods for automatic evaluation of delayed enhancement-mri. the results of the emidec challenge," *Medical Image Analysis*, vol. 79, p. 102428, 2022.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[7] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, pp. 424–432.

[8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[9] M. Z. Alom, C. Yakopcic, M. Hasan, T. M. Taha, and V. K. Asari, "Recurrent residual u-net for medical image segmentation," *Journal of Medical Imaging*, vol. 6, no. 1, p. 014006, 2019.

[10] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes," *IEEE transactions on medical imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[11] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz *et al.*, "Attention u-net: Learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

5. https://flare22.grand-challenge.org

6. https://github.com/HiLab-git/SSL4MIS

[12] Y. Zhang, L. Yuan, Y. Wang, and J. Zhang, "Sau-net: efficient 3d spine mri segmentation using inter-slice attention," in *Medical Imaging With Deep Learning*. PMLR, 2020, pp. 903–913.

[13] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[14] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[15] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2021, pp. 171–180.

[16] J. Ma, "Cutting-edge 3d medical image segmentation methods in 2020: Are happy families all alike?" *arXiv preprint arXiv:2101.00232*, 2021.

[17] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, 2020.

[18] Y. Zhang, Q. Liao, L. Ding, and J. Zhang, "Bridging 2d and 3d segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5 d solutions," *Computerized Medical Imaging and Graphics*, p. 102088, 2022.

[19] Q. Yao, L. Xiao, P. Liu, and S. K. Zhou, "Label-free segmentation of covid-19 lesions in lung ct," *IEEE transactions on medical imaging*, vol. 40, no. 10, pp. 2808–2819, 2021.

[20] L. Zhang, X. Wang, D. Yang, T. Sanford, S. Harmon, B. Turkbey, B. J. Wood, H. Roth, A. Myronenko, D. Xu *et al.*, "Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation," *IEEE transactions on medical imaging*, vol. 39, no. 7, pp. 2531–2540, 2020.

[21] Y. Zhang, Q. Liao, L. Yuan, H. Zhu, J. Xing, and J. Zhang, "Exploiting shared knowledge from non-covid lesions for annotation-efficient covid-19 ct lung infection segmentation," *IEEE journal of biomedical and health informatics*, vol. 25, no. 11, pp. 4152–4162, 2021.

[22] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on challenges in representation learning, ICML*, vol. 3, no. 2, 2013, p. 896.

[23] X. Wang, Y. Yuan, D. Guo, X. Huang, Y. Cui, M. Xia, Z. Wang, C. Bai, and S. Chen, "Ssa-net: Spatial self-attention network for covid-19 pneumonia infection segmentation with semi-supervised few-shot learning," *Medical Image Analysis*, vol. 79, p. 102459, 2022.

[24] H. Yao, X. Hu, and X. Li, "Enhancing pseudo label quality for semi-supervised domain-generalized medical image segmentation," *arXiv preprint arXiv:2201.08657*, 2022.

[25] C. Li, L. Dong, Q. Dou, F. Lin, K. Zhang, Z. Feng, W. Si, X. Deng, Z. Deng, and P.-A. Heng, "Self-ensembling co-training framework for semi-supervised covid-19 ct segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 11, pp. 4140–4151, 2021.

[26] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2274–2282, 2012.

[27] B. H. Thompson, G. Di Caterina, and J. P. Voisey, "Pseudo-label refinement using superpixels for semi-supervised brain tumour segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[28] Y. Shi, J. Zhang, T. Ling, J. Lu, Y. Zheng, Q. Yu, L. Qi, and Y. Gao, "Inconsistency-aware uncertainty estimation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 3, pp. 608–620, 2021.

[29] Z. Zhang, C. Tian, H. X. Bai, Z. Jiao, and X. Tian, "Discriminative error prediction network for semi-supervised colon gland segmentation," *Medical Image Analysis*, vol. 79, p. 102458, 2022.

[30] R. Wang, Y. Wu, H. Chen, L. Wang, and D. Meng, "Neighbor matching for semi-supervised learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 439–449.

[31] C. M. Seibold, S. Reiß, J. Kleesiek, and R. Stiefelhagen, "Reference-guided pseudo-label generation for medical semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 2, 2022, pp. 2171–2179.

[32] K. Han, L. Liu, Y. Song, Y. Liu, C. Qiu, Y. Tang, Q. Teng, and Z. Liu, "An effective semi-supervised approach for liver ct image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[33] Y. Ouali, C. Hudelot, and M. Tami, "An overview of deep semi-supervised learning," *ArXiv*, vol. abs/2006.05278, 2020.

[34] W. Huang, C. Chen, Z. Xiong, Y. Zhang, X. Chen, X. Sun, and F. Wu, "Semi-supervised neuron segmentation via reinforced consistency learning," *IEEE Transactions on Medical Imaging*, 2022.

[35] X. Li, L. Yu, H. Chen, C.-W. Fu, L. Xing, and P.-A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 2, pp. 523–534, 2020.

[36] G. Bortsova, F. Dubost, L. Hogeweg, I. Katramados, and M. d. Bruijne, "Semi-supervised medical image segmentation via learning consistency under transformations," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 810–818.

[37] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *Advances in neural information processing systems*, vol. 32, 2019.

[38] Y. Shu, H. Li, B. Xiao, X. Bi, and W. Li, "Cross-mix monitoring for medical image segmentation with limited supervision," *IEEE Transactions on Multimedia*, 2022.

[39] H. Basak, R. Bhattacharya, R. Hussain, and A. Chatterjee, "An embarrassingly simple consistency regularization method for semi-supervised medical image segmentation," *arXiv preprint arXiv:2202.00677*, 2022.

[40] K. Zheng, J. Xu, and J. Wei, "Double noise mean teacher self-ensembling model for semi-supervised tumor segmentation," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 1446–1450.

[41] M.-C. Xu, Y.-K. Zhou, C. Jin, S. B. Blumberg, F. J. Wilson, M. deGroot, D. C. Alexander, N. P. Oxtoby, and J. Jacob, "Learning morphological feature perturbations for calibrated semi-supervised segmentation," *arXiv preprint arXiv:2203.10196*, 2022.

[42] W. Luo, Y. Li, R. Urtasun, and R. S. Zemel, "Understanding the effective receptive field in deep convolutional neural networks," in *NIPS*, 2016.

[43] M. Xu, N. P. Oxtoby, D. C. Alexander, and J. Jacob, "Learning to pay attention to mistakes," *ArXiv*, vol. abs/2007.15131, 2020.

[44] Y. Li, L. Luo, H. Lin, H. Chen, and P.-A. Heng, "Dual-consistency semi-supervised learning with uncertainty quantification for covid-19 lesion segmentation from ct images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 199–209.

[45] Y. Ouali, C. Hudelot, and M. Tami, "Semi-supervised semantic segmentation with cross-consistency training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 12 674–12 684.

[46] X. Xu, T. Sanford, B. Turkbey, S. Xu, B. J. Wood, and P. Yan, "Shadow-consistent semi-supervised learning for prostate ultrasound segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1331–1345, 2021.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[48] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[49] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," *arXiv preprint arXiv:1610.02242*, 2016.

[50] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[51] X. Zeng, R. Huang, Y. Zhong, D. Sun, C. Han, D. Lin, D. Ni, and Y. Wang, "Reciprocal learning for semi-supervised segmentation," pp. 352–361, 2021.

[52] L. Yu, S. Wang, X. Li, C.-W. Fu, and P.-A. Heng, "Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation," in *International Conference on Medical Image Com-*

puting and Computer-Assisted Intervention. Springer, 2019, pp. 605–613.

[53] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" *Advances in neural information processing systems*, vol. 30, 2017.

[54] Z. Xie, E. Tu, H. Zheng, Y. Gu, and J. Yang, "Semi-supervised skin lesion segmentation with learning model confidence," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 1135–1139.

[55] X. Luo, G. Wang, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, D. N. Metaxas, and S. Zhang, "Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency," *Medical Image Analysis*, p. 102517, 2022.

[56] X. Luo, W. Liao, J. Chen, T. Song, Y. Chen, S. Zhang, N. Chen, G. Wang, and S. Zhang, "Efficient semi-supervised gross target volume of nasopharyngeal carcinoma segmentation via uncertainty rectified pyramid consistency," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 318–329.

[57] Z. Fang, J. Bai, X. Guo, X. Wang, F. Gao, H.-Y. Yang, B. Kong, Y. Hou, K. Cao, Q. Song *et al.*, "Annotation-efficient covid-19 pneumonia lesion segmentation using error-aware unified semi-supervised and active learning," *IEEE Transactions on Artificial Intelligence*, 2022.

[58] X. Zhao, C. Fang, D.-J. Fan, X. Lin, F. Gao, and G. Li, "Cross-level contrastive learning and consistency constraint for semi-supervised medical image segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[59] C. Chen, K. Zhou, Z. Wang, and R. Xiao, "Generative consistency for semi-supervised cerebrovascular segmentation from tof-mra," *IEEE Transactions on Medical Imaging*, 2022.

[60] A. Xu, S. Wang, S. Ye, J. Fan, X. Shi, and X. Xia, "Ca-mt: A self-ensembling model for semi-supervised cardiac segmentation with elliptical descriptor based contour-aware," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[61] G. Chen, J. Ru, Y. Zhou, I. Rekik, Z. Pan, X. Liu, Y. Lin, B. Lu, and J. Shi, "Mtans: Multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation," *NeuroImage*, vol. 244, p. 118568, 2021.

[62] L. Hu, J. Li, X. Peng, J. Xiao, B. Zhan, C. Zu, X. Wu, J. Zhou, and Y. Wang, "Semi-supervised npc segmentation with uncertainty and attention guided consistency," *Knowledge-Based Systems*, vol. 239, p. 108021, 2022.

[63] Y. Xie, J. Zhang, Z. Liao, J. Verjans, C. Shen, and Y. Xia, "Intra-and inter-pair consistency for semi-supervised gland segmentation," *IEEE Transactions on Image Processing*, vol. 31, pp. 894–905, 2021.

[64] Y. Liu, W. Wang, G. Luo, K. Wang, and S. Li, "A contrastive consistency semi-supervised left atrium segmentation model," *Computerized Medical Imaging and Graphics*, p. 102092, 2022.

[65] Z. Xu, Y. Wang, D. Lu, L. Yu, J. Yan, J. Luo, K. Ma, Y. Zheng, and R. K.-y. Tong, "All-around real label supervision: Cyclic prototype consistency learning for semi-supervised medical image segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[66] S. Li, C. Zhang, and X. He, "Shape-aware semi-supervised 3d semantic segmentation for medical images," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 552–561.

[67] S. Liu, Y. Li, X. Li, and G. Cao, "Shape-aware multi-task learning for semi-supervised 3d medical image segmentation," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1418–1423.

[68] X. Luo, J. Chen, T. Song, and G. Wang, "Semi-supervised medical image segmentation through dual-task consistency," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8801–8809.

[69] Y. Zhang and J. Zhang, "Dual-task mutual learning for semi-supervised medical image segmentation," in *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer, 2021, pp. 548–559.

[70] Y. Zhang, Q. Liao, R. Jiao, and J. Zhang, "Uncertainty-guided mutual consistency learning for semi-supervised medical image segmentation," *arXiv preprint arXiv:2112.02508*, 2021.

[71] Q.-Q. Chen, Z.-H. Sun, C.-F. Wei, E. Q. Wu, and D. Ming, "Semi-supervised 3d medical image segmentation based on dual-task consistent joint learning and task-level regularization," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[72] K. Wang, B. Zhan, C. Zu, X. Wu, J. Zhou, L. Zhou, and Y. Wang, "Tripled-uncertainty guided mean teacher model for semi-supervised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 450–460.

[73] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proceedings of the eleventh annual conference on Computational learning theory*, 1998, pp. 92–100.

[74] L. Zhu, K. Yang, M. Zhang, L. L. Chan, T. K. Ng, and B. C. Ooi, "Semi-supervised unpaired multi-modal learning for label-efficient medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 394–404.

[75] X. Chen, H.-Y. Zhou, F. Liu, J. Guo, L. Wang, and Y. Yu, "Mass: Modality-collaborative semi-supervised segmentation by exploiting cross-modal consistency from unpaired ct and mri images," *Medical Image Analysis*, p. 102506, 2022.

[76] X. Liu, F. Xing, N. Shusharina, R. Lim, C. Kuo, G. E. Fakhri, and J. Woo, "Act: Semi-supervised domain-adaptive medical image segmentation with asymmetric co-training," *arXiv preprint arXiv:2206.02288*, 2022.

[77] M. Liu, L. Xiao, H. Jiang, and Q. He, "Ccat-net: A novel transformer based semi-supervised framework for covid-19 lung lesion segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[78] Z. Xiao, Y. Su, Z. Deng, and W. Zhang, "Efficient combination of cnn and transformer for dual-teacher uncertainty-aware guided semi-supervised medical image segmentation," *Available at SSRN 4081789*.

[79] X. Luo, M. Hu, T. Song, G. Wang, and S. Zhang, "Semi-supervised medical image segmentation via cross teaching between cnn and transformer," in *Medical Imaging With Deep Learning*. PMLR, 2022, pp. 1–14.

[80] S. Qiao, W. Shen, Z. Zhang, B. Wang, and A. Yuille, "Deep co-training for semi-supervised image recognition," in *Proceedings of the european conference on computer vision (eccv)*, 2018, pp. 135–152.

[81] J. Peng, G. Estrada, M. Pedersoli, and C. Desrosiers, "Deep co-training for semi-supervised image segmentation," *Pattern Recognition*, vol. 107, p. 107269, 2020.

[82] Y. Xia, F. Liu, D. Yang, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, and H. Roth, "3d semi-supervised learning with uncertainty-aware multi-view co-training," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2020, pp. 3646–3655.

[83] Z. Zhao, J. Hu, Z. Zeng, X. Yang, P. Qian, B. Veeravalli, and C. Guan, "Mmgl: Multi-scale multi-view global-local contrastive learning for semi-supervised cardiac image segmentation," *arXiv preprint arXiv:2207.01883*, 2022.

[84] W. Dong-DongChen and Z.-H. WeiGao, "Tri-net for semi-supervised deep learning," in *Proceedings of twenty-seventh international joint conference on artificial intelligence*, 2018, pp. 2014–2020.

[85] P. Wang, J. Peng, M. Pedersoli, Y. Zhou, C. Zhang, and C. Desrosiers, "Self-paced and self-consistent co-training for semi-supervised image segmentation," *Medical Image Analysis*, vol. 73, p. 102146, 2021.

[86] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. IEEE, 2016, pp. 565–571.

[87] Y. Wang, Y. Zhang, J. Tian, C. Zhong, Z. Shi, Y. Zhang, and Z. He, "Double-uncertainty weighted method for semi-supervised learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 542–551.

[88] W. Hang, W. Feng, S. Liang, L. Yu, Q. Wang, K.-S. Choi, and J. Qin, "Local and global structure-aware entropy regularized mean teacher model for 3d left atrium segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 562–571.

[89] X. Hao, S. Gao, L. Sheng, and J. Zhang, "Parameter decoupling strategy for semi-supervised 3d left atrium segmentation," in *Fourteenth International Conference on Machine Vision (ICMV 2021)*, vol. 12084. SPIE, 2022, pp. 118–124.

[90] S. Li, Z. Zhao, K. Xu, Z. Zeng, and C. Guan, "Hierarchical consistency regularized mean teacher for semi-supervised 3d left atrium segmentation," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021, pp. 3395–3398.

[91] Y. Wu, M. Xu, Z. Ge, J. Cai, and L. Zhang, "Semi-supervised left atrium segmentation with mutual consistency training," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 297–306.

[92] C. You, Y. Zhou, R. Zhao, L. Staib, and J. S. Duncan, "Simcvd: Simple contrastive voxel-wise representation distillation for semi-supervised medical image segmentation," *IEEE Transactions on Medical Imaging*, 2022.

[93] Y. Zhang, L. Yang, J. Chen, M. Fredericksen, D. P. Hughes, and D. Z. Chen, "Deep adversarial networks for biomedical image segmentation utilizing unannotated images," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2017, pp. 408–416.

[94] H. Yang, C. Shan, A. Bouwman, L. R. Dekker, A. F. Kolen, and P. H. de With, "Medical instrument segmentation in 3d us by hybrid constrained semi-supervised learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 2, pp. 762–773, 2021.

[95] H. Peiris, Z. Chen, G. Egan, and M. Harandi, "Duo-segnet: adversarial dual-views for semi-supervised medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 428–438.

[96] H. Wu, G. Chen, Z. Wen, and J. Qin, "Collaborative and adversarial learning of focused and dispersive representations for semi-supervised polyp segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3489–3498.

[97] S. Li, Y. Zhang, and X. Yang, "Semi-supervised cardiac mri segmentation based on generative adversarial network and variational auto-encoder," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1402–1405.

[98] D. Nie, Y. Gao, L. Wang, and D. Shen, "Asdnet: attention based semi-supervised deep networks for medical image segmentation," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2018, pp. 370–378.

[99] J. Hou, X. Ding, and J. D. Deng, "Semi-supervised semantic segmentation of vessel images using leaking perturbations," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 2625–2634.

[100] K. Chaitanya, N. Karani, C. F. Baumgartner, A. Becker, O. Donati, and E. Konukoglu, "Semi-supervised and task-driven data augmentation," in *International conference on information processing in medical imaging*. Springer, 2019, pp. 29–41.

[101] H. Wu, Z. Wang, Y. Song, L. Yang, and J. Qin, "Cross-patch dense contrastive learning for semi-supervised segmentation of cellular nuclei in histopathologic images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 666–11 675.

[102] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," *Advances in neural information processing systems*, vol. 17, 2004.

[103] J. Wu, H. Fan, X. Zhang, S. Lin, and Z. Li, "Semi-supervised semantic segmentation via entropy minimization," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[104] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Mutual exclusivity loss for semi-supervised deep learning," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 1908–1912.

[105] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C. A. Raffel, E. D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," *Advances in neural information processing systems*, vol. 33, pp. 596–608, 2020.

[106] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," *Advances in neural information processing systems*, vol. 31, 2018.

[107] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: a regularization method for supervised and semi-supervised learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 8, pp. 1979–1993, 2018.

[108] H. Zheng, L. Lin, H. Hu, Q. Zhang, Q. Chen, Y. Iwamoto, X. Han, Y.-W. Chen, R. Tong, and J. Wu, "Semi-supervised segmentation of liver using adversarial learning with deep atlas prior," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2019, pp. 148–156.

[109] Y. He, G. Yang, J. Yang, Y. Chen, Y. Kong, J. Wu, L. Tang, X. Zhu, J.-L. Dillenseger, P. Shao *et al.*, "Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation," *Medical image analysis*, vol. 63, p. 101722, 2020.

[110] X. Hu, D. Zeng, X. Xu, and Y. Shi, "Semi-supervised contrastive learning for label-efficient medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 481–490.

[111] J. Wang, X. Li, Y. Han, J. Qin, L. Wang, and Z. Qichao, "Separated contrastive learning for organ-at-risk and gross-tumor-volume segmentation with limited annotation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 2459–2467.

[112] J. Yang, Y. Tao, Q. Xu, Y. Zhang, X. Ma, S. Yuan, and Q. Chen, "Self-supervised sequence recovery for semi-supervised retinal layer segmentation," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[113] R. Zhang, S. Liu, Y. Yu, and G. Li, "Self-supervised correction learning for semi-supervised biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 134–144.

[114] J. Peng, P. Wang, C. Desrosiers, and M. Pedersoli, "Self-paced contrastive learning for semi-supervised medical image segmentation with meta-labels," *Advances in Neural Information Processing Systems*, vol. 34, pp. 16 686–16 699, 2021.

[115] C. Dong, Y.-w. Chen, A. H. Foruzan, L. Lin, X.-h. Han, T. Tateyama, X. Wu, G. Xu, and H. Jiang, "Segmentation of liver and spleen based on computational anatomy models," *Computers in biology and medicine*, vol. 67, pp. 146–160, 2015.

[116] H. Park, P. H. Bland, and C. R. Meyer, "Construction of an abdominal probabilistic atlas and its application in segmentation," *IEEE Transactions on medical imaging*, vol. 22, no. 4, pp. 483–492, 2003.

[117] H. Huang, Q. Chen, L. Lin, M. Cai, Q. W. Zhang, Y. Iwamoto, X. Han, A. Furukawa, S. Kanasaki, Y. W. Chen *et al.*, "Mtl-abs3net: Atlas-based semi-supervised organ segmentation network with multi-task learning for medical images," *IEEE Journal of Biomedical and Health Informatics*, 2022.

[118] Z. Li, Z. Li, R. Liu, Z. Luo, and X. Fan, "Coupling deep deformable registration with contextual refinement for semi-supervised medical image segmentation," in *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2022, pp. 1–5.

[119] N. Zhang, J. Hou, R.-W. Zhao, R. Feng, and Y. Zhang, "Semi-supervised medical image segmentation with distribution calibration and non-local semantic constraint," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2021, pp. 1171–1178.

[120] J. Chen, H. Zhang, R. Mohiaddin, T. Wong, D. Firmin, J. Keegan, and G. Yang, "Adaptive hierarchical dual consistency for semi-supervised left atrium segmentation on cross-domain data," *IEEE Transactions on Medical Imaging*, vol. 41, no. 2, pp. 420–433, 2021.

[121] D. Kiyasseh, A. Swiston, R. Chen, and A. Chen, "Segmentation of left atrial mr images via self-supervised semi-supervised meta-learning," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 13–24.

[122] G. Zhang, Z. Yang, B. Huo, S. Chai, and S. Jiang, "Automatic segmentation of organs at risk and tumors in ct images of lung cancer from partially labelled datasets with a semi-supervised conditional nnu-net," *Computer methods and programs in biomedicine*, vol. 211, p. 106419, 2021.

[123] S. Reiß, C. Seibold, A. Freytag, E. Rodner, and R. Stiefelhagen, "Every annotation counts: Multi-label deep supervision for medical image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 9532–9542.

[124] Y. Lin, H. Yao, Z. Li, G. Zheng, and X. Li, "Calibrating label distribution for class-imbalanced barely-supervised knee segmentation," *arXiv preprint arXiv:2205.03644*, 2022.

[125] J. Wang and T. Lukasiewicz, "Rethinking bayesian deep learning methods for semi-supervised volumetric medical image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 182–190.

[126] Z. Xiong, Q. Xia, Z. Hu, N. Huang, C. Bian, Y. Zheng, S. Vesal, N. Ravikumar, A. Maier, X. Yang *et al.*, "A global benchmark of algorithms for segmenting the left atrium from late gadolinium-enhanced cardiac magnetic resonance imaging," *Medical Image Analysis*, vol. 67, p. 101832, 2021.

[127] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, "The cancer imaging archive (tcia): maintaining and operating a public information repository," *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013.

[128] Y. Zhou, Z. Li, S. Bai, C. Wang, X. Chen, M. Han, E. Fishman, and A. L. Yuille, "Prior-aware neural network for partially-supervised multi-organ segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 10 672–10 681.

[129] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[130] R. A. Zeineldin, M. E. Karar, J. Coburger, C. R. Wirtz, and O. Burgert, "Deepseg: deep neural network framework for automatic brain tumor segmentation using magnetic resonance flair images," *International journal of computer assisted radiology and surgery*, vol. 15, no. 6, pp. 909–920, 2020.

[131] O. Bernard, A. Lalande, C. Zotti, F. Cervenansky, X. Yang, P.-A. Heng, I. Cetin, K. Lekadir, O. Camara, M. A. G. Ballester *et al.*, "Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved?" *IEEE transactions on medical imaging*, vol. 37, no. 11, pp. 2514–2525, 2018.

[132] V. M. Campello, P. Gkontra, C. Izquierdo, C. Martin-Isla, A. Sojoudi, P. M. Full, K. Maier-Hein, Y. Zhang, Z. He, J. Ma *et al.*, "Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3543–3554, 2021.

[133] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang, "Abdomenct-1k: Is abdominal organ segmentation a solved problem," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[134] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[135] M. Van Waerebeke, G. Lodygensky, and J. Dolz, "On the pitfalls of entropy-based uncertainty for multi-class semi-supervised segmentation," *arXiv preprint arXiv:2203.03587*, 2022.

[136] Z. Ren, R. Yeh, and A. Schwing, "Not all unlabeled data are equal: Learning to weight data in semi-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 786–21 797, 2020.

[137] A. Ghosh and A. H. Thiery, "On data-augmentation and consistency-based semi-supervised learning," in *International Conference on Learning Representations*, 2020.

[138] K. Zhang and X. Zhuang, "Cyclemix: A holistic strategy for medical image segmentation from scribble supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 656–11 665.

[139] Z. Xu, D. Lu, Y. Wang, J. Luo, J. Jayender, K. Ma, Y. Zheng, and X. Li, "Noisy labels are treasure: mean-teacher-assisted confident learning for hepatic vessel segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2021, pp. 3–13.