Embracing Imperfect Datasets: A Review of Deep Learning Solutions for Medical Image Segmentation

Nima Tajbakhsh, Laura Jeyaseelan, Qian Li, Jeffrey Chiang, Zhihao Wu, and Xiaowei Ding[a]

[a] *VoxelCloud, Inc.*

Abstract

The medical imaging literature has witnessed remarkable progress in high-performing segmentation models based on convolutional neural networks. Despite the new performance highs, the recent advanced segmentation models still require large, representative, and high quality annotated datasets. However, rarely do we have a perfect training dataset, particularly in the field of medical imaging, where data and annotations are both expensive to acquire. Recently, a large body of research has studied the problem of medical image segmentation with imperfect datasets, tackling two major dataset limitations: scarce annotations where only limited annotated data is available for training, and weak annotations where the training data has only sparse annotations, noisy annotations, or image-level annotations. In this article, we provide a detailed review of the solutions above, summarizing both the technical novelties and empirical results. We further compare the benefits and requirements of the surveyed methodologies and provide our recommended solutions to the problems of scarce and weak annotations. We hope this review increases the community awareness of the techniques to handle imperfect datasets.

Keywords: medical image segmentation, imperfect dataset, scarce annotations, noisy annotations, unreliable annotations, sparse annotations, and weak annotations

1. Introduction

Medical imaging literature has witnessed great progress in the designs and performance of deep convolutional models for medical image segmentation. Since the introduction of UNet Ronneberger et al. (2015), neural architectures for medical image segmentation have transformed markedly. State-of-the-art architectures now benefit from re-designed skip connections Zhou et al. (2018b), residual convolution blocks Alom et al. (2018), dense convolution blocks Li et al. (2018), attention mechanisms Oktay et al. (2018), hybrid squeeze-excitation modules Roy et al. (2018), to name a few. Although the architectural advancements have enabled new performance highs, they still require large, high-quality annotated datasets—more so than before.

However, rarely do we have a perfectly-sized and carefully-labeled dataset to train an image segmentation model, particularly for medical imaging applications, where both data and annotations are expensive to acquire. The common limitations of medical image segmentation datasets include scarce annotations where only limited annotated data is available for training, and weak annotations where the training data has only sparse annotations, noisy annotations, or image-level annotations. In the presence of these dataset shortcomings, even the most advanced segmentation models may fail to generalize to datasets from real-world clinical settings. In response to this challenge, researchers from the medical imaging community have actively sought solutions, resulting in a diverse and effective set of techniques with demonstrated capabilities in handling scarce and weak annotations for the task of medical image segmentation. In this article, we have reviewed these solutions in depth, summarizing both the technical novelties and empirical results. We hope this review increases the community awareness of the existing solutions for the common limitations of medical image segmentation datasets, and further inspires the research community to explore solutions for the less explored dataset problems.

2. Related works

The recent methodological advancements in medical image analysis have been covered in several surveys. Yi et al. (2018) broadly investigated the use of GANs in medical imaging. Cheplygina et al. (2019) reviewed semi-supervised, multi-instance learning, and transfer learning in medical image analysis, covering both deep learning and traditional segmentation methods. Hesamian et al. (2019) surveyed deep learning techniques suggested for medical image segmentation but with a particular focus on architectural advancements and training schemes. Very recently, Zhang et al. (2019b) reviewed the solutions that tackle the small sample size problem for the broad medical image analysis.

In contrast, the current survey has focused on medical image segmentation rather than the broad medical image analysis, covering strategies for handling both scarce and weak annotations. The specific scope and deep review of this survey distinguish it from Yi et al. (2018); Cheplygina et al. (2019) that broadly cover deep learning for general medical image analysis, from Hesamian et al. (2019) that focuses on

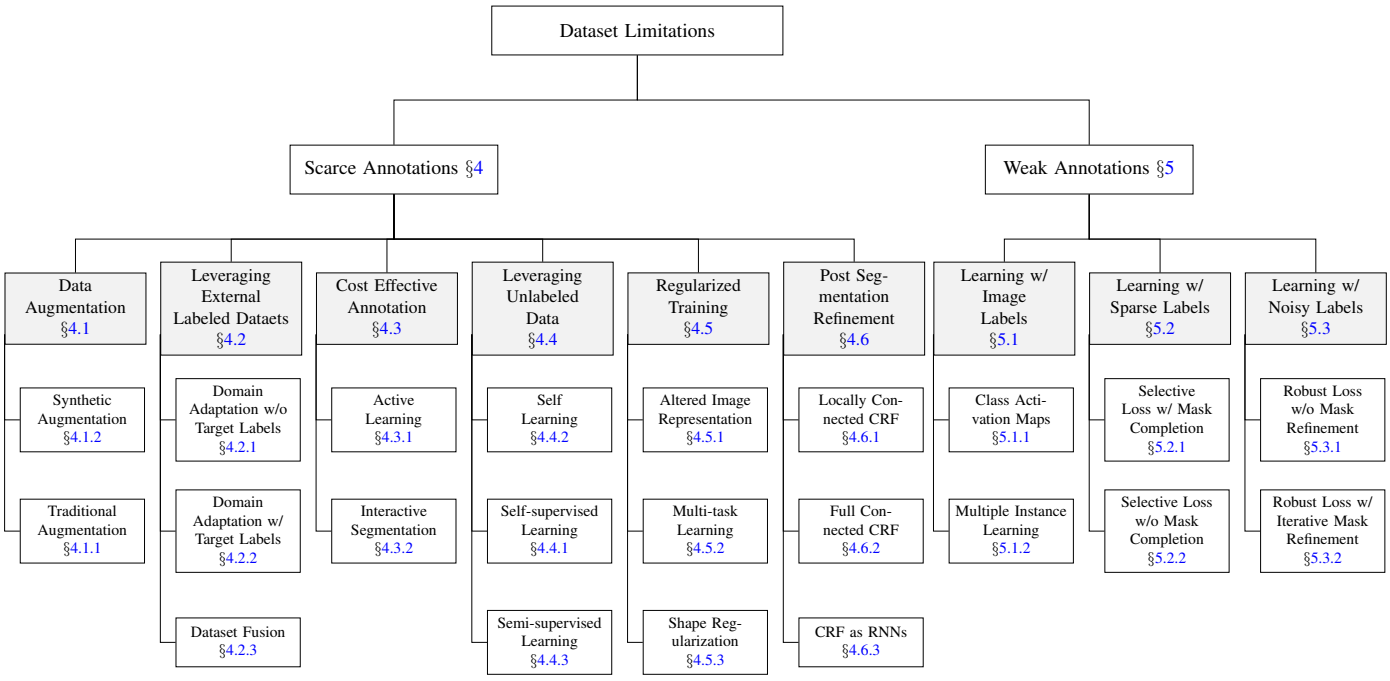


Figure 1: Organization of this review paper. We broadly categorize the limitations of medical image segmentation datasets into scarce annotations and weak annotations. For each problem, we then present the strategies (highlighted in grey) followed by the suggested solutions.

architectural advancements for medical image segmentation, and from Zhang et al. (2019b) that investigates only the small sample size problem in medical image segmentation.

3. Organization of survey

Figure 1 shows the common limitations of medical image segmentation datasets and the suggested solutions accordingly. We have broadly categorized dataset limitations into two categories: 1) scarce annotations (Section 4) where only a small fraction of images in the dataset are densely annotated; 2) weak annotations (Section 5) where the provided annotations are sparse, noisy, or only at image-level. We have further grouped the solutions for scarce annotations by their methodological principles, covering methods based on data augmentation in Section 4.1, leveraging external labeled datasets in Section 4.2, cost effective annotation in Section 4.3, leveraging unlabeled data in Section 4.4, regularized training in Section 4.5, and finally post segmentation refinement in Section 4.6. The solutions suggested for weak annotations are grouped by the type of weakness these methodologies tackle. Specifically, we cover methods for image-level annotations in Section 5.1, sparse annotations in Section 5.2, and noisy annotations in Section 5.3. We compare the solutions under review in Section 6, and provide our recommended solutions based on the underlying advantages and required resources. Finally, this survey is concluded in Section 7.

4. Problem I: Scarce annotation

Scarce annotation is a common problem when using supervised deep learning methods for medical image segmentation. Traditional solutions to this problem are data augmentation, transfer learning from natural images, and weight regularization. However, these techniques can only partially address the problem of limited annotation. For example, traditional data augmentation is handicapped by the large correlation between the original training set and the augmented examples. Transfer learning from natural images only benefits 2D medical image segmentation models, with no benefits to the common 3D medical image segmentation models.

The limited capability of the traditional methods in handling the problem of scarce annotations has led to the development of modern reactive and proactive approaches. The reactive methods tackle the problem of scarce annotation through a post segmentation refinement using variants of conditional random fields. The proactive approaches, on the other hand, actively enlarge the training set through cost-effective annotation and synthetic data generation or change the training paradigm by leveraging unlabeled data and using additional model regularization during training. In the following, we provide a comprehensive summary of such modern solutions to the ubiquitous problem of scarce annotations in medical image segmentation.

4.1. Data augmentation

Data augmentation has served as an effective solution to the problem of over-fitting, particularly in the absence of large labeled training sets. However, traditional data augmentation methods result in images that are highly correlated; as such, their impact can be limited. Apart from intrinsic high correlation between augmented and original images, rare conditions may also not be properly enhanced during typical

data augmentations. Synthetic data augmentation, on the other hand, samples more diverse examples, with sufficiently different visual appearance, from the same manifold as the original training set. In this section, we review both traditional and synthetic data augmentation methods as they can offer complimentary values.

4.1.1. Traditional Augmentation

Traditional data augmentation has proved effective in reducing over-fitting and improving test performance for both natural and medical images Zhang et al. (2016). The data augmentation methods used in medical imaging can be grouped by the image property they intend to manipulate Zhang et al. (2019a). These common image properties consists of image quality, image appearance, and image layout.

By image quality: Similar to the data augmentation for 2D natural images, image quality can be affected by sharpness, blurriness and noise. Christ et al. (2016) apply Gaussian noise to CT scans as part of data augmentation. Sirinukunwattana et al. (2017) employ Gaussian blur on colon histology images for the task of gland segmentation. Zhang et al. (2019a) show that data augmentation by adjusting image quality enables the largest performance gain in MR images, with largest improvement coming from image sharpening through the application of unsharp masking.

By image appearance: Data augmentation by adjusting image appearance consists in manipulating the statistical characteristics of the image intensities such as brightness, saturation and contrast. Liskowski and Krawiec (2016) apply gamma correction of saturation and value of the HSV color space prior to segmenting retinal blood vessels. Dong et al. (2017) employ random enhancement of brightness in 3D MR volumes to enrich the training set for brain tumor segmentation. Contrast augmentation is usually helpful when images exhibit inhomogeneous intensities. For instance, Fu et al. (2017) apply a contrast transformation function on fluorescence microscopy images to enrich the dataset for the task of nuclei segmentation. Alex et al. (2017) use histogram matching as a form of pre-processing where the 3D MR images are matched against an arbitrarily chosen reference image from the training data.

By image layout: Data augmentation by changing image layout consists of spatial transformations such as rotation, scaling and deformation. Ronneberger et al. (2015) show that augmenting the training set with random elastic deformations is key to training a segmentation network with very few annotated images. Milletari et al. (2016) also apply a dense deformation field through a 2x2x2 grid of control-points and B-spline interpolation on the training images. Çiçek et al. (2016) first sample random vectors from a normal distribution in a grid with a spacing of 32 voxels in each direction and then apply a B-spline interpolation.

4.1.2. Synthetic Augmentation

Synthetic data augmentation methods for medical image segmentation can be broadly grouped into same-domain and cross-domain image synthesis. The former consists of synthesizing labeled data directly in the target domain. The latter, on the other hand, consists of projecting labeled data from another domain to the target domain, which is closely related to the subject of domain adaptation. We therefore postpone a detailed review of cross-domain image synthesis methods until Section 4.2.2, where we present a detailed study of domain adaptation techniques. Nevertheless, we have summarized the representative approaches of both groups in Table 1. In the following, we review the methods suggested for same-domain image synthesis.

Guibas et al. (2017) propose a framework consisting of a GAN and a conditional GAN to generate pairs of synthetic fundus images and the corresponding vessel masks. Specifically, the GAN takes as input a random vector and then generates a synthesized vessel mask, which is then sent to the conditional GAN to generate the corresponding photo-realistic fundus image. The authors verify the fidelity of the synthesized images by examining whether a classifier can distinguish the synthetic images from the real images, but do not demonstrate whether the synthesized examples enable training a more accurate segmentation model.

Shin et al. (2018) use a conditional GAN to generate synthetic MR images given a lesion mask and a brain segmentation mask. Once trained, the synthesis network can generate synthesized MR images given a user-defined tumor mask. The elegance of this approach is in how the user can rescale or relocate a tumor in the mask and then the synthesis network can generate the MR image in accordance to the new size and location of the tumor. Without typical data augmentation, the tumor segmentation model trained using both synthetic and real MR images achieves a significant performance gain over the model trained using only real MR images. However, the performance gap is largely bridged in the presence of typical data augmentation. In a similar spirit, Mahapatra et al. (2018) use a conditional GAN to synthesize X-ray images with desired abnormalities. The model takes as inputs an X-ray with an abnormality and a lung segmentation mask, and then it generates a synthesized X-ray that has the same diseases as the input X-ray while taking the image appearance that matches the provided segmentation mask. This approach has the capability of generating many synthesized diseased images from one real diseased image.

Lung segmentation is challenging in the presence of large plueral nodules, which are often under-represented in the training sets. To overcome this limitation, Jin et al. (2018) train an image in-painting model based on a conditional GAN that can synthesize pleural nodules in the nodule-free CT slices. The authors demonstrate that the lung segmentation model trained using images with synthetic pleural nodules is superior to the model trained using only real images wherein the pleural nodules are under-represented.

Zhao et al. (2019) propose a data synthesis method to generate pairs of brain MR images and the segmentation masks from only one labeled MR image. For the task of brain structures segmentation, the suggested data augmentation method, which is further discussed in Section 4.4.3, enables four points increase in Dice over a model trained using

Table 1: Comparison between image synthesis methods suggested for medical image segmentation.

Publication	Synthesis Type	Domains	Description
Chartsias et al. (2017)	Cross-domain synthesis	CT $\rightarrow$ MRI	Cycle GAN is used to generate pairs of synthesized MR images from pairs of CT slices and the corresponding myocardium masks
Zhang et al. (2018c)	Cross-domain synthesis	CT $\leftrightarrow$MRI	Cycle GAN with shape consistency loss is used to translate between MR and CT scans. Segmentation and synthesis networks are trained jointly.
Guibas et al. (2017)	Same-domain synthesis	Fundus	GAN is used to generate a vessel mask and a conditional GAN is used to generate the corresponding fundus image
Shin et al. (2018)	Same-domain synthesis	MRI	Conditional GAN to generate synthetic MR images given a lesion mask and a brain segmentation mask
Jin et al. (2018)	Same-domain synthesis	CT	Conditional GAN is used to synthesize pleural nodules in the nodule-free CT slices
Zhao et al. (2019)	Same-domain synthesis	MRI	Hybrid spatial-intensity transformation network is used to synthesize MR images from 1 labeled MR image
Mahapatra et al. (2018)	Same-domain synthesis	X-ray	Conditional GAN is used to synthesize X-ray images with desired abnormalities

tradition data augmentation and 3 points increase in Dice over atlas-based data augmentation.

4.2. Leveraging External Labeled Datasets

A frequently encountered obstacle in medical imaging is that of a distribution shift between the data available for training and the data encountered in clinical practice. This shift could be caused by using different scanners and image acquisition protocols or due to imaging different patient populations and ethnicities. As individual datasets tend to be small and typically originate from a single institution, they are inherently biased and models trained on them perform poorly in the real world. Given the limitations of individual datasets, a natural workaround is to incorporate multiple datasets for training.

Annotations from external labeled datasets can be leveraged either through domain adaptation techniques or through dataset fusion. Domain adaptation techniques attempt to bridge the gap between multiple domains by either learning a latent representation that is common to these various domains or by learning to translate images from one domain to the other. These domains may consist of different imaging modalities or different image distributions within the same modality. Dataset fusion, on the other hand, simply utilizes data from different sources to train a general segmentation model having superior performance to those trained on each individual dataset. A bird's eye view of the papers discussed in this subsection is shown in Table 2

4.2.1. Domain Adaptation without Target Labels

When the test domain (a.k.a the target domain) labels are unavailable, but we only have access to labels from a different domain (a.k.a the source domain), the popular approach is to convert one domain to the other.

Source $\rightarrow$ Target: In the absence of target labels, one approach is to convert the source domain images to have the style of the target domain while retaining the anatomical structure and thereby the segmentation masks of the source domain. Then, a segmentation network trained on the target-styled images and source masks can be used to make predictions on the target images. Huo et al. (2018a) suggest a joint image synthesis and segmentation framework that enables image segmentation for the target domain using unlabeled target images and labeled images from a source domain. The intuition behind this joint optimization is that the training process can benefit from the complementary information between the synthesis and segmentation networks. In this framework, the main job is done by the image synthesis network, a CycleGAN, that converts the labeled source images to synthesized target images. During training, the synthesized target images are used to train the segmentation network. At test time, the real images from the target domain are directly submitted to the segmentation network to obtain the desired segmentation masks. The authors evaluate this framework for the task of spleen segmentation in CT scans where the CT scans do not have the segmentation masks, but the source MR images come with spleen masks. The authors show that the model trained using synthesized CT scans can achieve a performance level at par to the model trained using real CT scans with labels. A similar approach is taken by Huo et al. (2018b) for the task of splenomegaly and total intracranial volume segmentation, reporting 2% improvement in Dice over the existing state of the art—a 2-stage Cycle GAN followed by a separate segmentation network. For the task of total intracranial volume segmentation, the Dice coefficient using domain adaptation is only 1% lower than the Dice coefficient of the model trained with the target labels (upper bound). Chen et al. (2019) perform domain translation from the MR to CT domain for the task of heart CT segmentation using only MR image masks. They propose the use of a Cycle GAN for conversion from MR to CT and vice-versa with a segmentation network trained on the real and generated CT (target) images. The novelty of their approach lies in the use of a shared encoder common to both the CT segmentation network and the CT to MR generator network, which makes use of this multitask setting to prevent the segmentation encoder from over-fitting. The authors report a 9% improvement in Dice over the existing state of the art domain adaptation techniques.

Target $\rightarrow$ Source: Alternatively, one can convert the target domain images into the source domain followed by training the segmentation model using the source images. During inference, the target images are first converted to the source domain and then fed to the segmentation network to generate

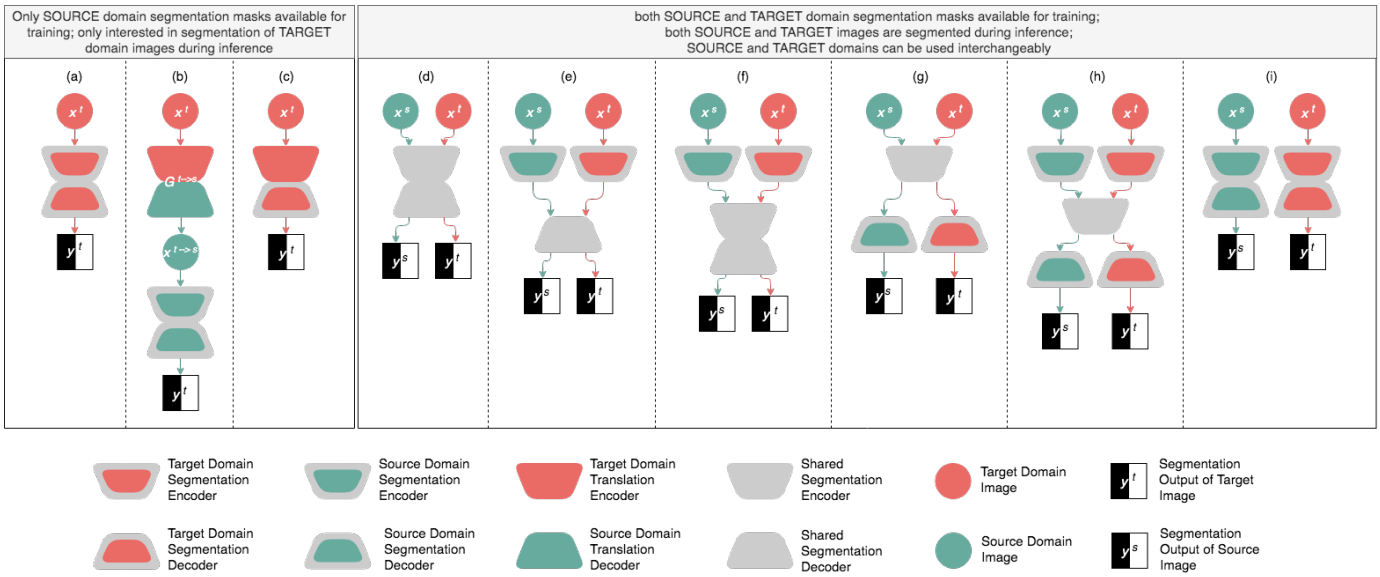


Figure 2: Leveraging external labeled datasets is effective for the problem of scarce annotations. This figure compares the data flow during inference for the related solutions: (a)-(c) cover approaches that only use source domain labels during training while (d)-(i) cover approaches that make use of both source and target domain labels, in which case the terms 'source' and 'target' are no longer meaningful and can be used interchangeably. (a) Target domain images are directly passed through the segmentation network trained on images from the target domain and images translated to the target domain. (b) Target domain images are converted to the source domain and then sent to the segmentation network trained on the source domain and source-like images. (c) Target domain images are sent through the target domain encoder (belonging to the domain translation network) and then sent to the segmentation decoder trained on target domain and target-like images. (d) Images from either domain can be passed through the segmentation network trained jointly on both domains. (e) Both domain images are passed through their own domain specific encoders and then through the segmentation decoder trained on both domains. (f) Similar to (e) but now the domain specific encoded feature maps are sent through a jointly trained segmentation encoder and decoder, (g) Images from both domains pass through a jointly trained segmentation encoder and then pass through domain specific decoders, (h) Each domain has its own specific encoder and segmentation decoder, but pass through a shared segmentation encoder in between. (i) Each domain has its own segmentation network during inference which is trained using data from its own domain augmented using domain translation.

the segmentation maps.

Giger (2018) propose converting the CT (target) domain to the MR (source) domain and then using an existing atlas-based algorithm (MALP-EM) to perform the segmentation on the converted MR images. The motivation is that it is easier to obtain segmentation annotations for Brain MRI than Brain CT scans. They use a modified U-Net for the domain conversion, which requires the CT and MR images to be registered beforehand. On average, they improve the Dice score by 9% over a baseline that performs segmentation in the CT domain.

Chen et al. (2018) use the cycle GAN with an additional semantic adversarial loss, which is used to distinguish between source segmentation masks and segmentation predictions of the converted target to source images. The authors evaluate their proposed method on 2 different X-ray datasets, which vary in disease type, intensity, and contrast. They achieve 2% improvement in Dice over the baseline cycle GAN performance.

Given a set of annotated CT scans, Zhang et al. (2018b) aim to segment X-ray images without having any X-ray segmentation annotations. For this purpose, the authors first convert annotated CT scans to digitally reconstructed radiographs (DRRs) via a 3D to 2D projection, and then learn a mapping between DRRs and X-ray images. The mapping is performed by a task-driven GAN, which is a Cycle GAN with an additional segmentation loss to generate segmentations for the DRR-style images. With these new constraints, the suggested method improves the segmentation Dice by two or three points over using either one of them alone and over the vanilla CycleGAN.

4.2.2. *Domain Adaptation with Target Labels*

If the segmentation masks are available for both domains, there is no longer a distinction between the choice of source and target domains. In this scenario, domain adaptation is achieved by learning a shared feature encoding, allowing the segmentation network to predict meaningful masks regardless of the input domain.

Chartsias et al. (2017) use cycle GAN to generate pairs of synthesized MR images and the corresponding myocardium masks from pairs of CT slices and their myocardium segmentation masks. The authors base the image synthesis module on CycleGAN, because it does not require the CT and MR images to be registered nor do they have to belong to the same patient. Once the synthetic data generated, the authors train a myocardium segmentation model using both synthetic MR and real MR images, demonstrating 15% improvement over the myocardium segmentation model trained using only the real MR images.

However, Zhang et al. (2018c) demonstrate that the above offline data augmentation may only be partially effective and in some cases can even deteriorate the performance. Instead, they propose a framework wherein both data synthesis model and segmentation model are trained jointly. They develop a

Table 2: Overview of the papers leveraging external labeled datasets. The suggested method, among other factors, differ in terms of presence of target domain labels and the domain in which segmentation is performed. The Figure column on the right shows the matching data flow from Figure 2.

Publication	Availability of Target Domain Segmentation Masks	Segmentation Domain	Modality	Figure
Domain Adaptation without Target Labels				
Huo et al. (2018a)	✖	Target	MRI, CT	(a)
Huo et al. (2018b)	✖	Target	MRI, CT	(a)
Chen et al. (2018)	✖	Source	X-ray	(b)
Zhang et al. (2018b)	✖	Source	DRR, X-ray	(b)
Chen et al. (2019)	✖	Target	MRI, CT	(c)
Giger (2018)	✖	Source	MRI, CT	(b)
Domain Adaptation with Target Labels				
Chartsias et al. (2017)	✔	Both	MRI, CT	(i)
Zhang et al. (2018c)	✔	Both	MRI, CT	(i)
Dou et al. (2018)	✔	Both	MRI, CT	(e)
Valindria et al. (2018)	✔	Both	MRI, CT	(d),(e),(f),(g),(h)
Dataset Fusion				
Harouni et al. (2018a)	✔	All domains	MRI,CT,US,X-ray	(d)
Dmitriev and Kaufman (2019)	✔	All domains	CT	(d)

segmentation network that can segment heart chambers in both CT and MR images by learning a translation between the two domains. They use a cycle GAN as their backbone and further add a shape consistency loss to ensure anatomical structure invariance during translation. They improve the segmentation Dice score on CT images by 8 points and MR images by 2 points over other methods that use both real and synthetic data for training.

Dou et al. (2018) train a cardiac segmentation network, consisting of two parallel domain-specific encoders and a shared decoder. During training, the decoder takes its input from a single encoder depending on the domain of the input image. The network is trained so that the decoder yields similar high-level semantic embedding for images of both domains. This is achieved by a discriminator that is trained to distinguish between the two domains. They achieve substantial performance boost over single domain training and 2% improvement in Dice over other domain adaptation techniques.

For the case where both source and target labels are available, domain adaptation is achieved using shared latent representations between the two domains, but the location of the shared features is a network design choice. Valindria et al. (2018) evaluate the performance of four different locations for the shared latent representations: 1) separate encoders with a shared decoder (see Figure 2(e)), 2) separate initial streams, followed by a shared encoder and decoder (see Figure 2(f)), 3) a shared encoder and separate decoder streams (see Figure 2(g)) and finally, 4) separate encoder and decoder streams with a shared latent representation in-between (see Figure 2(h)). They compared these variants with a baseline consisting of a single-stream encoder-decoder segmentation network, which is trained with data from both domains (see Figure 2(d)). Their results showed that the baseline was actually at par with or in some cases outperformed most variants, the only exception being the fourth variant, which consistently outperformed the baseline and other dual stream variants.

4.2.3. Dataset Fusion

Dataset fusion techniques leverage multiple datasets to train a universal segmentation model based on heterogeneous, disjoint datasets, offering two advantages: 1) more efficient training, as multiple models are consolidated into a single model, and 2) enhanced regularization, as data from multiple sources can provide further supervision. Domain adaptation and Dataset fusion both aim to leverage multiple datasets; however, they take different approaches: the former does this by minimizing the domain shift, whereas the latter does so by learning to discriminate between domains.

It is inefficient to have modality-specific models to segment the same organs across different modalities. Harouni et al. (2018a) propose a modality independent model that is jointly trained using data from all modalities. The network architecture is a modified U-Net with the base U-Net performing multi-organ segmentation and a classification head added to the bottleneck layer, which performs the modality/viewpoint classification (7 classes: X-ray, Short Axis MRI, 2chamber MRI, 4chamber MRI, CT, Ultra Sound 4chamber B-mode, Ultra Sound Doppler). The authors compared their jointly trained universal network against individually trained U-Nets for each task. The results show that the universal network usually performed at par with or outperformed the specialized networks. The exception to this performance was seen for left ventricle segmentation, where a dedicated MRI model showed significantly higher performance.

Dmitriev and Kaufman (2019) train a multi-organ segmentation model using data from multiple single organ datasets. For this purpose, the authors add an additional channel, which

is filled with a class-specific hash value, to each layer of the decoder network, conditioning the segmentation predictions on the class labels. The drawback, however, is that the test image with an unknown organ label must be fed in 'm' times sequentially to condition on all the possible classes. The multi-dataset training scheme achieves 1.5% improvement in Dice over the state of the art single dataset approaches.

4.3. Cost-effective Annotation

Perhaps, the most reliable approach to the scare annotation problem is to obtain additional labeled examples. This approach requires the availability of unlabeled medical images, access to a pool of expert annotators, and more importantly additional annotation budget. However, to fully utilize the annotation budget, one must decide how to choose examples for annotation from a large set of unlabeled images and how to accelerate the annotation process given the limited availability of medical experts. The former question is addressed by active learning methods, which determines the next batch of samples for annotation so as to maximize model's performance, and the latter is addressed by interactive segmentation, which assists the expert annotators by propagating their modifications through the entire segmentation mask.

4.3.1. Active Learning

Active learning is a cost-effective approach to enlarge the training datasets; and thus, it is highly amenable to the problem of limited annotation budget in medical image segmentation where clinical experts have limited availability, annotation cost is high, and the amount of unlabelled data is usually non-trivial. Active learning, in its general form, requires the availability of a base segmentation model; thus, a minimal set of base annotations is necessary. Therefore, datasets with no segmentation masks or those with only weak annotations may not directly benefit from active learning unless a pre-trained model from a similar domain is available to serve as the base segmentation model. In what follows, we present a high-level overview of the active learning paradigm and then review the active learning methods for medical image segmentation.

Active learning is an iterative paradigm wherein the unlabeled samples for each round of annotation are selected judiciously to maximally improve the performance of the current model. Algorithm 1 shows the pseudocode of active learning. In each iteration, the segmentation model is run against the unlabeled images, and then a set of selection criteria, which are defined on model outputs, are used to select the next batch of samples for annotation. Once annotated, the new batch is added to the training set and the segmentation model is fine-tuned using the augmented training set. This process is repeated until the performance on a validation set plateaus. Active learning methods differ in their sample selection criteria and their definition of annotation unit (the whole or only a part of the image is to be annotated). Table 3 compares the active learning methods suggested for medical image segmentation.

Yang et al. (2017) propose a framework called suggestive annotation where the candidate samples for each round of annotation are selected through a 2-stage screening process. First, uncertain samples are identified through the application of an ensemble of segmentation models. The uncertainty at pixel-level is computed as the variance of predictions generated by individual models in the ensemble. Pixel level uncertainty is then averaged to form one uncertainty value for the entire image. Second, the uncertain images are further refined by removing the samples that have high visual similarity. Suggestive annotation, with 50% of training data, achieves the same level of performance that can be obtained using the entire training dataset.

Algorithm 1: Active learning

Input : Initial model $\mathcal{M}_0$, unlabeled dataset $\mathcal{U}_0$, size of query batch k, iteration times $\mathcal{T}$, active learning algorithm $\mathcal{A}$

Output: Labeled dataset $\mathcal{L}_\mathcal{T}$, updated model $\mathcal{M}_\mathcal{T}$

$\mathcal{L}_0 \leftarrow \varnothing$;
for $i \leftarrow 1$ **to** $\mathcal{T}$ **do**
  /* phase 1: query batch selection */
  $\mathcal{Q}_t \leftarrow \mathcal{A}(\mathcal{U}_{t-1}, \mathcal{M}_{t-1}, k)$;
  annotate samples in $\mathcal{Q}_t$;
  /* phase 2: update model */
  $\mathcal{L}_t \leftarrow \mathcal{L}_{t-1} \cup \{(\mathbf{x}, y) | \mathbf{x} \in \mathcal{Q}_t, y \in \mathcal{Y}_t\}$;
  $\mathcal{M}_t \leftarrow$ fine-tuning $\mathcal{M}_{t-1}$ using $\mathcal{L}_t$;
  $\mathcal{U}_t \leftarrow \mathcal{U}_{t-1} \backslash \mathcal{Q}_t$;
end
return $\mathcal{L}_\mathcal{T}, \mathcal{M}_\mathcal{T}$

Kuo et al. (2018) propose an active learning framework based on sample uncertainty and annotation cost. In fact, this work is the first of its kind in the context of medical image segmentation to account for annotation cost when selecting the samples for the next round of annotation. Without considering annotation cost, active learning frameworks treat the images equally, ignoring the fact that some images in practice incur substantially higher annotation cost due to the larger size or quantity of contained target structures (organs and abnormalities). Concretely, they formulate active learning as a knapsack 0-1 problem where the objective is to select a batch of samples for annotation so as to maximize the model uncertainty while keeping the annotation cost below a given threshold. To measure sample uncertainty, they propose to train FCNs at the patch-level rather than the image-level because a patchFCN is less likely to overfit to the global image context. To estimate annotation cost for each unlabeled image, they use a regression model where the predictor variables are the total perimeter and number of connected components in the segmentation mask. The suggested cost-sensitive approach, with only 20% of annotation cost, can achieve a performance at par with a model trained using the entire training set.

The methods suggested by Kuo et al. (2018); Yang et al. (2017), despite their differences, both employ an ensemble of FCNs to estimate sample uncertainty, which is slow and computationally expensive as one needs to incrementally train an

Publication	Type	Sample selection strategy			Annotation unit
		informativeness	diversity	annotation cost	
Gorriz et al. (2017)	iterative	✓			whole 2D image
Yang et al. (2017)	iterative	✓	✓		whole 2D image
Ozdemir et al. (2018)	iterative	✓	✓		whole 2D image
Kuo et al. (2018)	iterative	✓	✓	✓	whole 3D image
Sourati et al. (2018)	iterative	✓	✓		2D image patch
Mahapatra et al. (2018)	1-shot	✓			whole 2D image
Sourati et al. (2019)	iterative	✓	✓		2D image patch
Zheng et al. (2019)	iterative		✓		2D image patch

Table 3: Comparison between active learning methods for medical image segmentation. The suggested methods differ in terms of the definition of the annotation unit and the criteria by which these units are selected for the next round of annotation.

ensemble of segmentation models after each round of annotation. A more computationally efficient approach to quantifying model uncertainty is to run a given sample through the model several times with the dropout layers on. Pixel uncertainty is then estimated as the entropy of averaged probabilities over different classes. This efficient sample uncertainty estimation is used by Gorriz et al. (2017) to realize a cost-effective active learning framework. Specifically, they compute an uncertainly value for a given unlabelled image by first obtaining an uncertainty map using the aforementioned dropout-based method followed by reducing the map to a single value through a weighted averaging scheme where the weights come from a distance transform map over the segmentation result. The idea is to assign higher importance to the uncertain pixels that are located farther away from the object boundaries. Once uncertainty values are computed for all unlabelled images, at each round of active learning, they select samples with high uncertainties as well as a batch of random samples for annotation. In each round, they also directly add samples that have the lowest levels of uncertainty along with their predicted masks to the training set. The rationale is that if the sample uncertainty is low, then the model has probably created a high-quality segmentation mask, which can be used for training without any further corrections. With this approach, the authors have demonstrated a 55% reduction in the annotation cost.

Similar to Yang et al. (2017), Ozdemir et al. (2018) propose a 2-stage active learning framework where stage 1 identifies uncertain examples whereas stage 2 selects the representative examples among the uncertain examples. The suggested method is however different in how uncertainty and representativeness are measured. The authors use the dropout-based approach to estimate an uncertainty map for each unlabelled image. To identify representative examples, they use the latent space learned by the segmentation network; however, to increase the discrimination power of the latent space, they train the segmentation network using an entropy-based regularization technique, which encourages diversity among the features of the latent space. The farther an uncertain sample is located from other examples in the latent space, the more representative the example. The uncertainty and representiveness metrics are further fused using Borda count. By ranking the examples using the fused metric metrics, the authors achieve similar performance to the model trained with the full dataset while using only 27% of the entire training set.

Sourati et al. (2018) propose a probabilistic active learning framework where the probability of an unlabeled sample being queried in the next round of annotation is estimated based on its Fisher information. A sample has higher Fisher information if it generates larger gradients with respect to the model parameters. To incorporate Fisher information in the sample selection process, the authors formulate active learning as an optimization problem where the unknowns are the probabilities by which unlabelled samples are queried for the next round of annotation; the constraints are that the querying probabilities should add up to one and that they should change disproportionately to their Fisher information; and the objective is to assign the querying probabilities so as to maximize the overall Fisher information. The optimization problem above is solved for a batch of samples, as such, the sample inter-dependency is already taken into consideration, eliminating the need for a secondary stage that further selects the representative samples from the informative samples. This one-shot behaviour sets this approach apart from the previous works where informativeness and representativeness are accounted for sequentially (e.g., Yang et al. (2017)). One limitation of this work, however, is that computational complexity is super quadratic with respect to the number of parameters, because the Fisher matrix has as many rows and columns as the number of parameters in the network. This limitation has been addressed in a follow-up work from the authors (Sourati et al. (2019)) where the number of rows and columns of the Fisher matrix reduces to the number of layers in the network. This scheme, when trained with only 0.5% of the training voxels, can achieve the same performance as the model that is trained using the entire training set.

Mahapatra et al. (2018) use a Bayesian neural network for active learning where the informative samples are selected using a combined metric based on aleotaric uncertainty (noise in the data) and epistemic uncertainty (uncertainty over the CNN parameters). Thus, this sample selection strategy differs from the previous approaches, where user-defined heuristics such as standard deviation of predictions are used to identify informative samples for annotation. Combined with an image

synthesis network, the suggested method achieves the full-data performance with only 30-35% of annotated pixels.

Different from previous works, Zheng et al. (2019) propose a 1-shot active learning method, which eliminates the need for iterative sample selection and annotation. The suggested method consists of a feature extraction network, which projects each image patch to a latent space, and a clustering algorithm, which discovers representative images for image annotation in the latent space. Being a 1-shot active learning approach, the feature extraction network must be trained using unlabeled data. For this purpose, the authors use various unsupervised models such as auto-encoders and variational auto-encoders. For clustering, the authors use a hybrid method based on K-means and max-cover algorithms. The results for both 2D and 3D datasets suggest that the 1-shot active learning method performs comparably to an iterative alternative by Yang et al. (2017).

4.3.2. *Interactive Segmentation*

Creating segmentation masks is not only tedious and time-consuming for expert annotators, but also incurs substantial annotation cost particularly for volumetric medical images where the same lesion or organ must be delineated across multiple slices. Interactive segmentation can accelerate the annotation process by allowing the expert annotators to interactively correct an initial segmentation mask generated by a model. Interactive segmentation complements active learning in achieving cost-effective annotation: the latter identifies which images to be annotated whereas the former reduces the time required to complete the annotation of a selected image.

Algorithm 2 shows the pseudocode for interactive segmentation. As seen, interactive segmentation may require an initial segmentation model, whose output is reviewed by human experts to provide feedback on possible segmentation error. The user feedback, as the core part of interactive segmentation methodologies, can take varying forms of interactions such as mouse clicks, bounding boxes, and scribbles. The user interactions then translate to foreground or background annotations, which the initial segmentation model can use to improve itself. The updated model re-generates the segmentation mask for the users' feedback, and this process repeats until the desired segmentation mask is obtained. Interactive segmentation is highly effective to cope with a model's inevitable segmentation mistakes, which are typically caused by domain shifts or unrepresentative training sets. In what follows, we summarize the recent interactive segmentation methods that are suggested for medical image segmentation.

Sun et al. (2018a) propose an interactive method for segmenting fuzzy boundaries, wherein the user first places a point roughly at the center of the object, and then the model performs object delineation for the user-specified structure. For accurate boundary segmentation, the authors suggest a segmentation model that delineates the structures by comparing the appearances of image patches from inside and outside of the structure, imitating inside-outside comparison that physicians perform in order to precisely localize boundaries. The

Algorithm 2: Interactive segmentation

Input : Initial model $\mathcal{M}_0$, unlabeled image $\mathcal{I}$, number of iterations $\mathcal{N}$, feedback operation $\mathcal{R}$, conversion operation $\mathcal{C}$

Output: Updated model $\mathcal{M}_\mathcal{N}$

for $i \leftarrow 1$ **to** $\mathcal{N}$ **do**
  `/* generate segmentation map */`
  $\mathcal{S}_i \leftarrow \mathcal{M}_{i-1}(\mathcal{I})$;
  `/* get feedback from an expert */`
  $\mathcal{F}_i \leftarrow \mathcal{R}(\mathcal{S}_i, \mathcal{I})$;
  `/* convert to a new annotation */`
  $\mathcal{A}_i \leftarrow \mathcal{C}(\mathcal{F}_i)$;
  $\mathcal{M}_i \leftarrow$ fine-tuning $\mathcal{M}_{i-1}$ with $\mathcal{A}_i$;
end
return $\mathcal{S}_\mathcal{N}$

authors model the inside-outside comparison with a bidirectional convolutional recurrent neural network, which is trained using the image patch and ground truth mask sequences bidirectionally, allowing the network to learn appearance changes from foreground to background and vice versa. This method, however, only allows users to specify a seed point at the onset of segmentation, that is, the resulting segmentation masks are not responsive to the subsequent user interaction.

Wang et al. (2018b) proposes a framework consisting of a proposal network and a refinement network where the former generates a base segmentation mask whereas the latter refines the base mask according to the suggestions provided by the user. However, the suggested framework lacks adaptability to unseen image contexts. The authors have overcome this limitation in their followup work Wang et al. (2018a). Given a test image and a pre-trained segmentation model, the suggested framework alternates between 2 steps: 1) refining the current segmentation mask through the application of Graph Cut Boykov and Jolly (2001), and 2) minimizing segmentation loss for the test image by creating a pseudo ground truth segmentation mask. This approach can be viewed as a self-learning method with the difference being the pseudo ground truth depends on both model predictions and user-provided scribbles. Specifically, the pseudo ground truth mask is the predicted segmentation mask wherein the labels of the scribble pixels are overwritten by the labels provided by the user. The segmentation loss is then a weighted cross entropy function, which receives large contributions from the scribble pixels and zero contributions from uncertain pixels. The authors treat a pixel as uncertain if it is located near a scribble but has a predicted label other than that of the scribble or if the posterior distribution predicted by the model has high entropy (low confidence predictions). The suggested framework outperforms traditional interactive segmentation methods in both accuracy and speed of annotation.

Sakinis et al. (2019) propose a semi-automated image segmentation method that enables a high quality segmentation with only a few user clicks. The authors choose a mouse click as the means of user interaction because it enables quick

feedback and ease of simulation. The segmentation model is a U-Net that receives as input the image stacked with the foreground and background attention maps, where attention maps are constructed by placing a Gaussian blob at each foreground and background user click. The U-Net is then trained by minimizing the Dice loss between the predicted segmentation and ground truth. Since it is unfeasible to have true user interaction during training and large-scale testing, the authors propose a simulation scheme that has the effect of a hypothetical user clicking on regions with larger and more noticeable segmentation error. This method proves effective in segmenting both structures that exist in the training set and the structures that the model has never seen during training. To put this in perspective, with only 1 user click, this semi-automated method can achieve a Dice of 0.64 for segmenting colon cancer, outperforming the best automated model with a Dice of 0.56.

4.4. Leveraging Unlabeled Data

Unlabeled medical images, although lack annotations, can still be used in conjunction with labeled data to train higher-performing segmentation models. We have identified three scenarios wherein unlabeled medical images have aided medical image segmentation: 1) self-supervised pre-training where unlabeled images are used to pre-train a segmentation network; 2) self-learning where unlabeled images are labeled by a segmentation model and then used as new examples during training; and 3) semi-supervised learning where both labeled and unlabeled images are used jointly to train a segmentation model.

4.4.1. Self-supervised Pre-training

Transfer learning has commonly been used to tackle the limited training sample size problem in medical imaging, where models pre-trained on ImageNet are fine-tuned for target medical image analysis tasks. Despite promising results Tajbakhsh et al. (2016); Hoo-Chang et al. (2016), this approach has two major limitations. First, it may limit the designer to architectures that have been pre-trained on ImageNet, which are often needlessly deep for medical imaging, thus retarding training and inference. Second, transfer learning is barely applicable to 3D medical image analysis applications, because the 2D and 3D kernels are not shape compatible. Therefore, transfer learning from natural images is only a partial solution to the common problem of insufficient labeled data in medical imaging.

Recently, self-supervised model pre-training has been studied as an alternative solution to the problem of limited annotations in medical imaging. The key idea consists of pre-training the model using unlabeled medical data, which is easier to obtain, and then fine-tune the pre-trained model for the target medical vision task using the limited labeled data available for training. Specifically, self-supervised pre-training consists of assigning surrogate or proxy labels to the unlabeled data and then training a randomly initialized network using the resulting surrogate supervision signal. The advantage of model pre-training using unlabeled medical data is that the learned knowledge is related to the target medical task; and thus, can be more effective than transfer learning from a foregin domain (e.g., Tajbakhsh et al. (2019) and Ross et al. (2018)).

Self-supervised learning methods differ in the composition of the surrogate task. Reviewing the literature, we have identified two types of surrogate tasks: 1) image-to-scalar where an encoder network is pre-trained for a surrogate image classification or regression task; 2) image-to-image where an encoder-decoder network is pre-trained for a surrogate image regression task such as image colorization or image de-noising. While the former approach seems particularly suitable for a downstream image classification task, it can still be used to initialize the encoder of a segmentation network, in which case the decoder should be initialized with random weights. We have summarized the representative examples of both categories in Table 4, and further review them as follows.

Image-to-scalar: Jamaludin et al. (2017) propose longitudinal relationships between medical images as the surrogate task to pre-train model weights. To generate surrogate supervision, they assign a label of 1 if two longitudinal studies belong to the same patient and 0 otherwise. Zhang et al. (2017a) propose a surrogate task wherein two slices are randomly selected from a CT volume and then the encoder is to predict if one slice is above or below the reference slice. The pre-trained model is then fine-tuned for the task of body part recognition in CT and MR images. Tajbakhsh et al. (2019) use prediction of image orientation as the surrogate task where the input image is rotated or flipped and the network is trained to predict such a transformation. The authors show that this surrogate task is highly effective for diabetic retinopathy classification in fundus images and lung lobe segmentation in chest CT scans. Spitzer et al. (2018) propose a new surrogate task that can be used to pre-train a siamese network by predicting the 3D distance between two patches sampled from the same brain regions. The pre-trained model is then fine-tuned for cytoarchitectonic segmentation of human brain. Similarly, Gildenblat and Klaiman (2019) suggest a surrogate scheme to pre-train a siamese network by learning similarity between image patches. Specifically, the network is trained to distinguish between similar patches (nearby patches) and dissimilar patches (spatially distant patches). The pre-trained siamese network is then fine-tuned for tumor tile retrieval in histopathology images.

Image-to-image: Alex et al. (2017) use noise removal in small image patches as the surrogate task, wherein the surrogate supervision was created by mapping the patches with user-injected noise to the original clean image patches. Ross et al. (2018) use image colorization as the surrogate task, wherein color colonoscopy images are converted to gray-scale and then recovered using a conditional GAN. The pre-trained models are then fine-tuned for the task of instrument segmentation in colonoscopy videos. The instrument segmentation model pre-trained via self-supervised learning achieves a Dice score of 0.59, which outperforms the counterpart model pre-trained using Microsoft COCO dataset with a Dice score of 0.57. A similar study is also done by Tajbakhsh et al. (2019)

Table 4: Comparison between self-supervised training methods that can directly or indirectly aid medical image segmentation.

Publication	Network	Surrogate task		
		Type	Description	Annotation
Jamaludin et al. (2017)	Encoder	Image-to-scalar	Predict if two longitudinal studies belong to the same patient	1(same)/0(different)
Zhang et al. (2017a)	Encoder	Image-to-scalar	Predict the order of two slices random selected from the same CT scan	0(top)/1(bottom)
Tajbakhsh et al. (2019)	Encoder	Image-to-scalar	Predict the degree of rotation applied to a chest CT scan	$\frac{\theta}{90^\circ}$ ($\theta \in \{0, 90, 180, 270\}$)
Spitzer et al. (2018)	Siamese	Image-to-scalar	Predict the distance between two patches sampled from the same MR image	Float distance
Gildenblat and Klaiman (2019)	Siamese	Image-to-scalar	Predict if two patches sampled from the same MR image are spatially near	1(near)/0(far)
Alex et al. (2017)	Encoder-decoder	Image-to-image	Learn how to remove noise from MR image patches	Original patch before injecting noise
Ross et al. (2018)	Encoder-decoder	Image-to-image	Learn how to colorize gray-scale colonoscopy frames	Original frame before removing color
Tajbakhsh et al. (2019)	Encoder-decoder	Image-to-image	Learn how to colorize gray-scale tele-med skin images	Original image before removing color
Zhou et al. (2019)	Encoder-decoder	Image-to-image	Learn how to restore the image from various degradation transformations	Original image before degradation
Bai et al. (2019)	Encoder-decoder	Image-to-image	Learn how to weakly localize anatomical landmarks in MR images	Approximate landmark positions

wherein colorization is used as a surrogate task for skin segmentation in tele-medicine images. While image colorization proved more effective than training from scratch, it was outperformed by transfer learning from ImageNet, presumably because the distance between tele-medicine skin images and ImageNet is small. Bai et al. (2019) propose anatomical position prediction as a self-supervised scheme. The landmarks are however obtained through an annotation-free process based on the relative views of image planes. The cardiac MR segmentation model pre-trained with the surrogate task above achieves four points increase in Dice over a U-Net trained from scratch.

The self-supervised learning methods described above are limited to a specific surrogate scheme. Models Genesis suggested by Zhou et al. (2019) is a significant shift from this paradigm where a library of diverse self-supervised schemes, all formulated as an image restoration task, is used to generate self-supervision signal. The suggested framework is scalable to a large library of surrogate tasks, because all tasks in the library can share the same encoder and decoder during training, eliminating the need for task-specific decoders. The authors have evaluated Models Genesis for both image classification and segmentation in seven 2D and 3D medical datasets, demonstrating three and five points increase in IoU over 3D models trained from scratch for lung nodule segmentation and liver segmentation in CT scans, respectively.

4.4.2. *Self Learning*

Self-learning is an iterative framework wherein a model iteratively improves itself by learning from its own predictions on unlabeled data. A common assumption behind self-learning is that a labeled training set, even though small, is available for training. A more extreme yet less common scenario is where no initial labeled data is available for training. Self-learning has shown promising performance, producing models that outperform the counterparts trained using only labeled data.

Algorithm 3 shows the pseudo code for the common scenario in self-learning where one has access to both a small labeled dataset and a fairly large unlabeled dataset. First, a base model is trained using limited labeled data. The base model is then applied to unlabeled data to generate pseudo segmentation masks. The limited labeled data is then merged with pseudo-labeled data to update the base model. Self-learning alternates between the two steps above until a desired level of performance on the validation set is achieved. In a less common scenario, no initial labeled dataset is available for training, in which case, an unsupervised segmentation method such as K-means is used to generate pseudo masks for unlabeled data.

Algorithm 3: Self learning

Input : Small labeled dataset $\mathcal{L}$, unlabeled dataset $\mathcal{U}$, iteration times $\mathcal{T}$, masks generation function $\mathcal{F}$
Output: Updated model $\mathcal{M}_\mathcal{T}$

$\mathcal{M}_0 \leftarrow$ training base model with $\mathcal{L}$;
for $i \leftarrow 1$ **to** $\mathcal{T}$ **do**
  /* generate pseudo segmentation masks */
  $\mathcal{S}_i \leftarrow \mathcal{F}(\mathcal{M}_{i-1}, \mathcal{U})$;
  $\mathcal{D}_i \leftarrow \mathcal{L} \cup \{(\mathbf{x}, s) | \mathbf{x} \in \mathcal{U}, s \in \mathcal{S}_i\}$;
  $\mathcal{M}_i \leftarrow$ fine-tuning $\mathcal{M}_{i-1}$ using $\mathcal{D}_i$;
end
return $\mathcal{M}_\mathcal{T}$

Self-learning methods commonly follow the iterative process stated above, but they differ in how they initialize the base model, how they generate pseudo masks, and whether or not they use a mechanism to handle label noise in pseudo segmentation masks. We have compared the self-learning methods suggested for medical image segmentation from these perspectives in Table 5, and further review them as follows.

Without initially labeled dataset: Zhang et al. (2018a) train a cyst segmentation model using unlabeled chest CT scans. Since the dataset is completely unlabeled, the authors generate the initial ground truth using K-means clustering followed by a refinement stage through graph cuts. The segmentation model is trained using the pseudo masks and then the model is applied back to the data to generate refined pseudo masks. The training process alternates between updating the segmentation model and refining pseudo masks. In 3 iterations, the suggested method achieves 12-point increase in Dice over a model trained using the initial pseudo mask generated by K-means.

With initially labeled dataset: Bai et al. (2017) propose a two-step framework to segment the heart chambers in MR images. The training process alternates between two steps: 1) estimating the ground truth for unlabeled data using the current segmentation model followed by a refinement stage through the application of CRF, 2) updating the current model using both the labeled data with expert annotations and the

Table 5: Comparison between the self-learning methods for medical image segmentation. The suggested methods differ in how the initial labeled dataset is constructed, how pseudo annotations for unlabeled data are generated, and whether or not any special treatment is applied to the unreliable regions in pseudo annotation masks.

Publication	Initial annotations by	Pseudo masks generated by	Label noise handled by
Zhang et al. (2018a)	K-means	single segmentation model	N/A
Bai et al. (2017)	expert	single segmentation model + CRF	N/A
Zhou et al. (2018a)	expert	ensemble segmentation model	N/A
Min et al. (2018)	expert	ensemble segmentation model	a two-stream network
Nie et al. (2018)	expert	single segmentation model	a discriminator network

unlabeled data with pseudo annotations. This approach is simple to implement; however, hindered by the quality of pseudo annotations, the resulting model achieves only a moderate level of improvement over the model trained using only labeled data. Similarly, Zhou et al. (2018a) propose an iterative self-learning framework, but, at each iteration, the authors train three segmentation models for the axial, sagittal, and coronal planes. Once trained, the three models scan each unlabeled 3D image slice-by-slice, generating three segmentation volumes, which are further combined through a majority voting scheme to form the final segmentation mask. The unlabeled images with their estimated segmentation masks are added to the labeled set to train three new segmentation models in the next iteration. The authors evaluate their self-learning paradigm in segmenting 16 structures in abdominal CT scans, achieving on average four points increase in Dice over the model trained using only the labeled data.

A limitation with the previous approaches is that the images that have expert annotation and images with pseudo segmentation masks are treated equally during training. As such, errors in the pseudo labels can degrade the quality of the resulting models. To overcome this limitation, Min et al. (2018) propose a two stream network where each stream has its own independent weights. During training, if a training sample receives the same class prediction from both streams, then the sample is deemed as easy or hard, in which case it will not contribute to the overall loss, where the training samples refer to each individual pixel. The rationale is that the easy examples do not add much value to the model and the hard examples may have label noise; therefore, it is safe to exclude them from backpropagation. To obtain the pseudo segmentation masks, the authors propose a hybrid method based on model distillation Gupta et al. (2016) and data distillation Radosavovic et al. (2018), which essentially consists of model-ensembling and test-time data augmentation. The authors show that the suggested self-learning framework based on pseudo-labeling and the two-stream network is effective for myocardium and brain lesion segmentation in MR images, outperforming fully supervised models.

The inadequate quality of pseudo segmentation masks is further addressed in Nie et al. (2018) where the authors use only the reliable regions of the pseudo segmentation masks during training. Specifically, they propose a framework consisting of a segmentation network (generator) and a confidence network (discriminator), which are trained through an adversarial game. The discriminator, a fully convolutional network, serves two purposes: 1) distinguishing between ground truth and predicted masks at the pixel level, 2) providing a confidence (reliability) value for each pixel in the predicted mask. The former functionality is used during adversarial training whereas the latter functionality is used to identify reliable regions in the pseudo masks of unlabeled data. During training, the pseudo mask for an unlabeled image is first masked by the binarized confidence map and is then used as ground truth to compute the segmentation loss. The authors show the effectiveness of the suggested framework over a pure supervised approach across multiple datasets.

4.4.3. Semi-supervised Learning

Semi-supervised learning enables training of models using both labeled data and arbitrary amounts of unlabeled data. Unlike self-learning, majority of semi-supervised methods do not attempt to generate psuedo annotations for unlabeled data, rather, they use unlabeled data as is during training. The domain of unlabeled data may be the same as the labeled dataset, in which case semi-supervised learning can serve as an effective solution for the problem of limited annotations. This is because the model is presented with a larger number of samples during training; and thus, it may generalize better to the unseen test set. On the other hand, the unlabeled dataset may be from a domain other than that of the labeled dataset, in which case semi-supervised learning mitigates the domain shift problem by adapting the model to a target domain.

Algorithm 4 shows the pseudocode for semi-supervised learning in its most general form. As seen, the semi-supervised framework consists of two loss functions: a supervised loss function to which only labeled data contribute; and an unsupervised loss function or a regularization term, which is computed for both labeled and unlabeled data (see Table 6). The total loss is the summation of two terms, which is minimized for batches of labeled and unlabeled data. In what follows, we explain the semi-supervised methods suggested for medical image segmentation.

The semi-supervised framework suggested by Baur et al. (2017) consist of a U-Net with two loss functions: a Dice-based segmentation loss that is computed based on the labeled data; and an embedding loss, which, given a batch of labeled and unlabeled data, brings the feature embedding of

Algorithm 4: Semi-supervised learning

Input : Limited labeled dataset $\mathcal{L}$, unlabeled dataset $\mathcal{U}$, shared backbone $\mathcal{M}_c$, branch model and loss function for labeled data $\mathcal{M}_l, \ell_l$, branch model and loss function for unlabeled data $\mathcal{M}_u, \ell_u$

Output: Fine-tuned model $\mathcal{M}$

$\zeta_l \leftarrow \ell_l(\mathcal{M}_l(\mathcal{M}_c(\mathcal{L})))$;

$\zeta_u \leftarrow \ell_u(\mathcal{M}_u(\mathcal{M}_c(\mathcal{U})) + \ell_u(\mathcal{M}_u(\mathcal{M}_c(\mathcal{L}))$;

minimize$(\zeta_l + \zeta_u)$;

return $\mathcal{M}$

Table 6: Semi-supervised methods for medical image segmentation. The suggested methods combine the segmentation task with an unsupervised task, allowing the model to use both labeled and unlabeled images during training.

Publication	Unsupervised task
Bai et al. (2017)	Embedding consistency
Zhang et al. (2017b)	Image classification
Sedai et al. (2017)	Image reconstruction
Baur et al. (2017)	Manifold learning
Chartsias et al. (2018)	Image reconstruction
Huo et al. (2018a)	Image synthesis
Zhao et al. (2019)	Image registration
Li et al. (2019)	Transformation consistency

the same-class pixels as close as possible while pushing apart the feature embedding of the pixels from different classes. To identify same-class pixels between labeled and unlabeled images, the authors assume the availability of a noisy label prior for unlabeled images. Also, to reduce the number of pair-wise comparisons between feature embedding of all pixels within the batch, they employ a pixel sampling scheme. The suggested semi-supervised framework proves promising in improving the segmentation of multiple sclerosis in the presence of limited data and domain shift.

Sedai et al. (2017) propose a semi-supervised learning framework consisting of a segmentation network and an auto-encoder. The training process begins with training the reconstruction network, a variational auto-encoder, which stores the knowledge learned from the unlabeled images in its latent space. Next, the segmentation network, which is also a variational auto-encoder, is trained using the labeled data. To leverage the knowledge learned from the unlabeled data, in addition to the segmentation loss, the segmentation network benefits from an l2-loss between its latent feature vector and the one generated by the reconstruction network for a given labeled image. They evaluate this framework in segmenting optic cup segmentation in fundus images, demonstrating four points increase (under limited data) and one point increase (using full dataset) in Dice over the segmentation models that are trained using only labeled data.

Zhang et al. (2017b) propose a semi-supervised learning framework according to an adversarial game between a segmentation network (U-Net) and an evaluation network (encoder). Given an input image, the segmentation network generates a segmentation map, which is then stacked with the original image and fed to the evaluation network, resulting in a quality score. During training, the segmentation network is updated with two objectives: 1) minimizing the segmentation loss for the labeled images and 2) making the evaluation network assign a high quality score to the unlabeled images. On the other hand, the evaluation network is updated so as to assign a low quality score to unlabeled images but a high quality score to labeled images. Owing to this adversarial learning, the segmentation network enjoys a supervision signal from both labeled and unlabeled images. For the task of gland segmentation in histopathology images, the authors have demonstrated one point increase in Dice over fully-supervised models trained with labeled data.

Chartsias et al. (2018) propose a solution to the problem of domain shift based on a disentangled image representation where the idea is to separate information related to segmenting the structure of interest from the other image features that readily change from one domain to another. By doing so, the segmentation network focuses on the intrinsic features of the target structure rather than variations related to imaging scanners or artifacts. Being amenable to semi-supervised training, this solution is also effective for the problem of limited annotation. Specifically, this framework consists of a decomposer network and a reconstructor network. The decomposer network receives an input image and then generates a segmentation map and a latent vector, which are both later used by the reconstructor network to reconstruct the input image. To disentangle the information related to the target structure from the rest of the image, the authors binarize the predicted segmentation map, retaining only the shape information of the target structure. Facing only the shape information in the predicted mask, the reconstructor needs to find the remaining missing information from the latent vector in order to reconstruct the input image. As such, during training, the latent vector and predicted mask tend to accumulate complementary information about the image, leading to a disentangled image representation. In this framework, the segmentation loss, which is defined for the decomposer network, is trained using only the labeled data. To leverage unlabeled data, the reconstructor network is trained using both labeled and unlabeled data. The authors show that the suggested framework is highly effective for myocardial segmentation in low-data regime, but the performance gap closes as the size of training set increases.

Li et al. (2019) propose a semi-supervised learning framework consisting of a segmentation loss whose optimization requires labeled data, and a transformation-consistent regularization loss that is computed for both labeled and unlabeled data. Let x, T, S denote the input image, an image transformation, and the segmentation network, respectively. The transformation-consistent regularization is formulated as $\|S(T(x)) - T(S(x))\|$, which is essentially the mean squared error loss between the segmentation map of the transformed image, $S(T(x))$, and the transformed segmentation map of the input image, $T(S(x))$. Intuitively, this regularization term promotes the segmentation network to be equivariant with

respect to some user-defined transformations. For instance, if an image, being labeled or unlabeled, is rotated by 90 degrees, then the resulting segmentation map should also appear with 90 degrees of rotations with respect to the segmentation mask of the original image. Even though maintaining equivariance seems like an obvious and must-have characteristic, it is surprising how often segmentation networks fail to retain this property. By imposing this additional regularization term, not only does the model behave more predictably, but also one can include unlabeled data in the training process. The authors demonstrate that, given small quantities of training data, this semi-supervised learning framework is effective for skin lesion segmentation from dermoscopy images, optic disc segmentation from fundus images, and liver segmentation from volumetric CT scans.

Zhao et al. (2019) propose a semi-supervised learning framework for 1-shot medical image segmentation where only one example in the training set has the segmentation mask and the rest of images are unlabeled. For this purpose, the authors suggest a hybrid spatial-intensity transformation model. The spatial transformation network deforms the labeled image so it takes the spatial layout of a given unlabeled image. Once the layout is taken care of, the intensity transformation network changes the intensity at each pixel so the labeled image takes the appearance of a given unlabeled image. Together, the two transformation networks enable the generation of new labeled examples from a reference labeled image and a number of unlabeled images. For the task of brain structure segmentation in MR images, the segmentation model trained using the artificially labeled images achieves between three to four points increase in Dice over the models trained with traditional and atlas-based data augmentation.

Mondal et al. (2018) propose a semi-supervised framework that integrates labeled and unlabeled images for the task of brain tissue segmentation in MR images. For this purpose, the authors train the segmentation network by introducing an additional fake class, resulting in a $k+1$ class segmentation problem where k is the number of classes present in the dataset. During training, the segmentation network seeks to maximize the probability of the correct class for each pixel in the labeled images. For unlabeled images, the segmentation network minimizes the probability of each pixel belonging to the fake class, which has the effect of maximizing the probability by which an unlabeled pixel belongs to one of the k classes. The suggested method proves effective in segmenting brain tissues from MR images when the training set contains only a few annotated images.

4.5. Regularized Training

Having a large number of parameters, deep supervised models are prone to over-fitting, particularly in the absence of large training sets. The traditional regularization to the problem of over-fitting is weight regularization whereby the network is encouraged to keep the weights small, resulting in a simpler and more robust model. While effective, weight regularization is only one form of model regularization. In this section, we review other forms of regularization: altered image representation, which either creates a lower dimensional input space or changes the input space to ease the task of representation learning for the model; multi-task learning, which regularizes model weights through additional supervision signals; and shape regularization, which imposes a shape prior on the predicted segmentation results.

4.5.1. Altered Image Representation

Altered image representations consist of projecting or transforming the images into a more informative or compact representation, which present deep models with an easier problem to solve, thereby reducing the need for large training sets. Informative representations can be particularly effective for 2D medical image segmentation whereas compact representations can benefit 3D applications where the curse of dimensionality requires large annotated datasets. This section reviews altered representations for both 2D and 3D images.

Altered 3D image representation: Training segmentation models with altered 3D image representations include training 2D models with multi-scale and multi-view patches Wang et al. (2017), fusing 2D models trained for the three clinical views Xia et al. (2018), training a 2D model with a 2.5D image representation Angermann et al. (2019), and finally training a 3D model with a 3D representation augmented with handcrafted features Ghafoorian et al. (2017). We explain these methods in more detail as follows.

Wang et al. (2017) make use of multi scale 2D patch-based pixel predictions for the task of lung nodule segmentation. The network has three shallow branches, one for each of the three orthogonal clinical views that share the same central pixel. Each branch is fed a 2D 2-channel input that captures the nodule at two different scales. The three branches are then fused to provide a binary prediction for the central pixel of the patch. Thus, there are six new 2D patches used for every voxel in the 3D volume being segmented. The segmentation results for all voxels are finally put together to obtain the 3D segmentation mask. They achieve 7% increase in Dice when compared against GrabCut. However, they have not compared their method against a single view CNN approach.

Xia et al. (2018) propose a 2-stage approach where the first stage uses a set of 2D segmentation networks whose outputs are further fused in the second stage through a 3D volumetric fusion network. The 2D networks generate slice-by-slice predictions along each of the 3 orthogonal views—axial, sagittal and coronal. The stacks of predicted segmentation masks for the 3 views are then concatenated with the original image, creating a 4-channel input to the second network, which learns to fuse the predictions to produce the final 3D prediction. The suggested method improves Dice score by 1% over baselines that use majority voting for volume fusion.

Angermann et al. (2019) make use of intensity projections, specifically maximum intensity projections (MIP) at multiple angles, which are then fused to create a 2.5D representation of (Magnetic Resonance Angiography) MRA images. However, the authors have not demonstrated a significant gain over the 2D and 3D performance baselines.

Ghafoorian et al. (2017) use an image representation based on registered T1 and FLAIR MR images augmented with dense handcrafted features to segment white matter hyperintensities. Their multi-scale architecture equipped with the hand-crafted features achieves 6% increase in Dice over their single scale baseline that does not incorporate the handcrafted features. Noteworthy, the improved segmentation performance is mainly due to the contribution of handcrafted features rather than the multi-scale paradigm.

Altered 2D image representation: The aforementioned methods offered altered representations for 3D images; however, even problems that are inherently 2-dimensional can benefit from using a different representation. Fu et al. (2018) utilize an image representation based on the polar coordinate system for the purpose of joint cup and disc segmentation in fundus images. Specifically, the authors use a circular image crop around the cup and disc region, which is then converted to a rectangular image through a transformation from Cartesian to polar coordinates. Scale-based data augmentation is performed by varying the radius of the circle prior to coordinate conversion. The authors have used the area under the ROC curve for evaluating the binary glaucoma classification performance using the cup to disc ratio. The suggested image representation enables a 4% gain in AUC compared with the existing state of the art trained using the standard Cartesian plane image representation.

4.5.2. Multi-task Learning

Multi-task learning Zhang and Yang (2017) refers to the paradigm in which multiple tasks are derived from a single learned representation. In modern applications, this can be realized by a single feature extractor (encoder) on which multiple tasks (e.g. classification, detection, segmentation) are performed. Intuitively, this paradigm encourages the encoder network to learn a latent representation that generalizes across the required tasks, with each task serving as a regularizer for the others. The studies outlined below demonstrate that adding a parallel task generally results in improved segmentation performance. These tasks may be supervised (e.g. classification, detection), or unsupervised (e.g. image reconstruction).

Most multi-task learning applications to medical image segmentation involve a variant of the U-net. The upsampling (segmentation) branch can be understood as just one of multiple output "heads" connected to a feature extractor, which allows for a natural extension to other tasks from the abstract feature space. Mehta et al. (2018) apply such a multi-task U-net to segment different tissue types in breast biopsy histopathology images, where an additional classification head is trained to classify whether the image is malignant or not. Simply adding this additional branch, which encouraged the model to learn a feature space relevant to diagnosis in addition to tissue segmentation, significantly improves the IoU of the vanilla network by 7%.

Similarly, Jaeger et al. (2018); Huang et al. (2018) propose a joint segmentation and detection framework. Huang et al. (2018) use their method for colorectal tumor segmentation in MRI volumes. Their proposed model resembles Mask-RCNN He et al. (2017), in which a global image encoder network detects regions of interest (ROIs), and a local decoder performs tumor segmentation on the proposed regions. The feature pyramid representing each ROI is passed to the local decoder, which, unlike Mask-RCNN, only segments the region of interest. The authors show that the local (rather than global) decoding approach preserves spatial details as well as decreases GPU footprint. This approach outperforms other U-net variants by several points and the popular Mask R-CNN algorithm by 20 points when the Dice score is used for performance evaluation.

Sun et al. (2018b) apply multi-task learning to brain tissue segmentation, combining the segmentation task with image de-noising. The base architecture of their model can still be described as a modified U-net, in which a shared encoder generates a latent feature representation for two separate decoders (for image reconstruction and segmentation). However, instead of training the network from scratch, Sun and colleagues pre-train the reconstruction head before training the full model end-to-end. By jointly performing image reconstruction with segmentation, they observed improved segmentation performance in all types of brain tissue across different metrics.

One drawback of the above approaches is that the companion task (classification or detection) requires additional labels for each image. Image reconstruction can be considered an unsupervised task that still provides the regularization benefits of multi-task learning. Myronenko (2018) propose a framework combining image reconstruction and segmentation for the task of brain lesion segmentation. The proposed model is a variational auto-encoder with an asymmetrically large encoder backbone and two decoders: the first decoder is trained to reconstruct the input MR image whereas the second decoder generates teh segmentation maps. The suggested method outperforms other modifications such as CRF post-processing as well as sophisticated test-time data augmentation, winning the first place in the 2018 BRaTS challenge.

In addition to providing robustness to the learned feature representation via regularization, the multi-task paradigm provides a feasible framework for consolidated biomedical image segmentation. As Harouni and colleagues Harouni et al. (2018b) note, given the sheer number of potential conditions, a single model per condition is not a scalable approach for clinical use. As a proof of concept, Harouni and colleagues use an architecture similar to Y-net Mehta et al. (2018) to segment different organs in several different imaging modalities using a single network. A single U-net is trained to segment nine different targets encountered in thoracic imaging, which is complemented by a classification branch trained to determine the input domain (e.g., CT, MRI, Ultrasound). While this model does not exceed the state of the art performance for any one given modality or target, performance is competitive across all domains. The authors also report that the presence of the image classification branch results in slightly improved segmentation performance.

4.5.3. Shape Regularization

Shape defines a region of interest (ROI) in segmentation problems under certain constraints, e.g. smooth and semantically sound. Such constraints can be effectively encoded as regularization towards more realistic appearance of the segmentation output, especially when well annotated data is scarce. Specifically, shape regularization consists of imposing a prior, highlighting certain geometric and structural characteristics, on the segmented ROIs, by operating either at pixel-level with an emphasis on shape and boundary explicitly, or in depth to capture high level features related to semantic meanings. In this section, we refer to the methods serving the former and latter objective as *Shallow regularization* and *Deep regularization*, respectively.

Shallow regularization

Shallow shape prior may regularize boundary pixels towards a certain class of shapes. Mirikharaji and Hamarneh (2018) leverage a star shape prior via an extra loss term on top of a binary cross entropy loss. To regularize a segmented ROI towards a star shape, any point on the linear path in between the ROI center and an interior point is expected to be interior as well (Fig 3 a & b), ensuring a smooth segmentation mask without holes. This definition comprises a broad class of objects even including convex shapes as a special case. With this additional term, authors report a 3.0% gain in Dice using U-net as segmentation network over the same network without star shape regularized loss.

Another class of shallow priors operate on boundary pixels to further improve segmentation accuracy around the boundaries. Boundary points, the first order derivative of a region, better capture the proximity of two shapes by inducing an extra penalty term between the estimated and expected pixels along the boundary. Inspired by the optimization technique for computing gradient flows of curve evolution, Kervadec et al. (2019) introduce a non-symmetric L_2 loss to regularize boundary deviation of the segmentation mask S relative to the ground truth G,

$$\text{Dist}(\partial G, \partial S) = \int_{\partial G} ||q_{\partial S}(p) - p||^2 \mathrm{d}p, \qquad (1)$$

where a boundary point p on ∂G (boundary of GT) is aligned against its counterpart q on ∂S (boundary of prediction), which is written as $q_{\partial S}(p)$, such that $p \rightarrow q$ is norm to GT boundary at point p (see Fig 3 c). By using boundary loss for the task of brain lesion segmentation in MR Images, the authors report an 8% gain in Dice and a 10% gain in Hausdorff score over a baseline that uses generalized Dice as the loss function.

Deep regularization

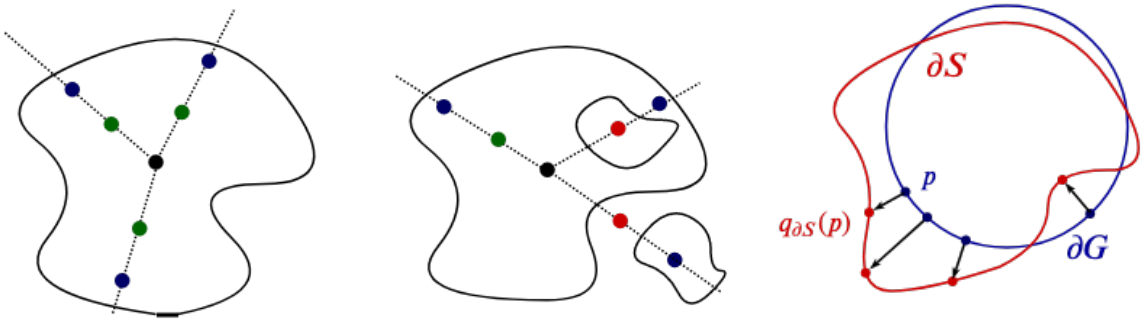
Shape regularization can also be applied to high level semantic features. Compared to shallow approaches, deep regularization, a.k.a deep supervision or deep priors, is less prone to image noise and more semantically and structurally aware. Ravishankar et al. (2017) incorporate deep prior within segmentation framework where a segmentation FCN is followed by a shape regularization FCN, which functions as a *convolutional de-noising autoencoder* (CDAE), consisting of an encoder that projects the segmentation mask to the shape space and a decoder that samples a segmentation mask from the shape space. In addition to the reconstruction loss, the regularization FCN has a projection loss that constraints ground truth and predicted segmentation to have similar encodings in the shape space. Combining both data augmentation and deep regularization, Ravishankar et al. (2017) report a 4.66% gain in Dice relative to a vanilla U-Net for the problem of ultrasound kidney segmentation.

Oktay et al. (2017) adopts a similar cascaded architecture with a major difference: the regularization FCN is first pre-trained as an auto-encoder with ground truth masks, and then only its frozen encoder is used as a regularizer during training the segmentation network. Therefore, the objective function reduces to a regular segmentation loss and a shape projection loss. The suggested model achieves 1.2% and 2.0% improvement in Dice over Ravishankar et al. (2017) for endocardium segmentation and myocardium segmentation, respectively. The authors attribute the inferior performance by Ravishankar et al. (2017) to over-regularization, which they have overcome br replacing CDAE with a frozen encoder during training.

Dalca et al. (2018) suggest a segmentation VAE that leverages shape prior in order to learn from unpaired images and segmentation masks. The VAE consists of an image encoder, which is initialized from scratch, and a frozen decoder, which is selected from an auto-encoder that has previously been trained for the task of mask reconstruction. Since the VAE uses a segmentation decoder, it generates a segmentation mask given an input MR image. However, the input MR images have no corresponding ground truth segmentation; therefore, the VAE is trained by minimizing the L2-loss between the input MR image and the predicted segmentation after being transformed through a 1x1 convolution block. During inference, the authors use the decoder output as the segmentation result. For the task of brain structure segmentation, the suggested method achieves Dice scores ranging between 0.50 and 0.80 on T1w images without any comparison against supervised methods.

4.6. Post segmentation refinement

Bias correction can be conducted as post hoc refinement regardless of the application domain. Among all post-processing techniques, Conditional Random Field (CRF) is the most commonly adopted and recognized approach to refine segmentation masks of both natural images (Schwing and Urtasun (2015), Chen et al. (2017) and Zheng et al. (2015)) and medical images (Roth et al. (2015), Chen and de Bruijne (2018) and Wachinger et al. (2018)). To obtain more realistic predicted masks, a CRF model incorporates two regularization terms: a smoothness term that removes small isolated regions, and an appearance term that ensures nearby pixels with similar color will more likely belong to the same class. The segmentation result, inferred as a maximum a posteriori (MAP) estimate from the CRF defined across all pixels, is expected to capture both local features and spatial dependency more holistically. As a consequence, CRF is able to refine a collection of



(a) (b) (c)

Figure 3: Shape regularization can combat the limited annotation problem by imposing additional constraints on predicted segmentation masks. Two common shape regularization methods are 1) star shape prior where any point in between the center and an interior point is constrained to be interior, and 2) boundary regularization. (a) An example of a segmented region that meets star shape prior. The black circle indicates the center of the segmented shape, the blue circles refer to interior points near the boundary, and the green circles are interior points that satisfy the star shape criteria.(b) An example of a segmented region with a hole and an isolated island, which does not meet requirements of star shape prior. The red circles indicate the pixels that lie exterior to the segmented region. (c) Boundary regularization improves segmentation accuracy around boundaries by minimizing point-wise deviation between the segmentation mask S relative to the ground truth G.

inaccurate and coarse pixel-level predictions, producing sharp boundaries and fine-grained segmentation masks.

Concretely, a CRF models pixel-wise labels, X, collectively as a random field that is conditioned upon image/volume intensities, I. This CRF can be characterized by its potentials, consisting of unary and pairwise terms Krähenbühl and Koltun (2011),

$$E(\mathbf{x} \mid \mathbf{I}) = \sum_i \phi_u(x_i) + \sum_{i \neq j} \phi_p(x_i, x_j). \quad (2)$$

The unary potential $\phi_u(x_i)$ is computed independently per pixel, which incorporates shape, texture, location and color descriptors. The existing variants of CRF differ in terms of the definition of the pairwise potential term, $\phi_p(x_i, x_j)$, and the underlying optimization technique. These variants include *Local CRF*, which considers neighboring pixels only, i.e. $j \in neighbor(i)$; *Fully Connected CRF* (FC-CRF), which considers all pixel pairs with an iterative mean field approximation of Eq. 2; and *RNN-CRF*, which takes a similar approach to FC-CRF, but it is now end-to-end trainable using Recurrent Neural Networks (RNNs). A visual comparison of the three types of CRF is provided in Fig 4.

4.6.1. Locally Connected CRF

Restricting the pairwise potentials to neighbouring pixels, the resulting CRF is designed to induce local smoothness. Roth et al. (2015) explore 2D CRF as well as a 3D Gaussian smoothing as post-processing for pancreas segmentation in CT images. The weights corresponding to pairwise and unary potentials are calibrated by a grid-search. In terms of performance, the authors report a 3.3% gain in Dice using CRF that falls short of a 6.9% gain using Gaussian smoothing. However, CRF does reduce the standard deviation of Dice in all experiments, demonstrating its regularization capability in reducing inference variance. Cai et al. (2016) use CRF to fuse mask and boundary predictions, which are separate branches off the same backbone during training, as a cascaded task that post-processes pancreas segmentation of MR Images. In this work, CRF still operates on neighbouring pixels in a feature space spanned by hand-crafted image features and the features learned by both segmentation branches. By using CRF for decision fusion, the authors report a 2.3% gain in Dice (73.8% pm 12.0% $\rightarrow$ 76.1% $\pm$ 8.7%) over a baseline without CRF.

4.6.2. Fully Connected CRF (FC-CRF)

Krähenbühl and Koltun (2011) provides an efficient inference approach to Eq. 2 using mean field approximation. The resulting algorithm reduces the computational complexity from quadratic to linear in the number of pixels involved in the computation. FC-CRF has proven effective as a segmentation post-processing solution for both natural images Chen et al. (2017) and 2D medical images, e.g. Fu et al. (2016b) in retinal images, Gao et al. (2016) on individual CT slices. The work by Kamnitsas et al. (2017) is the first to extend FC-CRF to 3D brain lesion segmentation in MR Images, leveraging intensity and spatial association under 3D context. However, their 3D generalization leads to marginal performance gains in Dice, i.e. 3.7% over Random Forests, 0.3% over an ensemble method, and merely 0.7% over their proposed architecture, which is a patch-based multi-scale 3D CNN network. In addition, the authors note that configuring 3D FC-CRF is a laborious task. Other 3D FC-CRF endeavors include a U-net + 3D FC-CRF by Christ et al. (2016) for liver and lesion segmentation in CT images and a 3D FC-CRF with spectral coordinates characterization by Wachinger et al. (2018) for neuroanatomy segmentation in MR Images. In the works above, FC-CRF refines segmentation masks that often exhibit small isolated regions and zigzag boundaries, but its effectiveness is greatly hindered by the extensive manual tweaking, or in other words, being not end-to-end trainable.

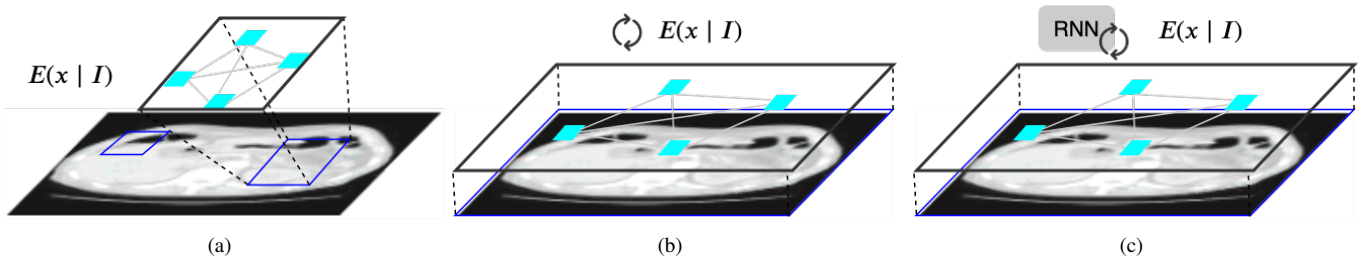


Figure 4: Post segmentation refinement can, to some degree, correct the segmentation errors. (a) local CRF optimizes Gibbs energy over local patches weighing in pairwise pixel dependency. (b) Fully Connected CRF (FC-CRF) extends the local scope of the CRF to the whole image in an efficient manner. (c) CRF as Recurrent Neural Networks (RNN-CRF) makes FC-CRF end-to-end trainable by replacing the iterative calculations with an RNN.

4.6.3. CRF as Recurrent Neural Networks (RNN-CRF)

RNN-CRF organically integrates CRF with CNNs, making it possible to train the whole network in an end-to-end manner. Zheng et al. (2015) reformulates the mean field approximation of FC-CRF as a stack of common CNN layers and the iterative optimization as hidden states in an RNN. Fu et al. (2016a) combine a multi-scale and multi-level CNN that has auxiliary output layers with a RNN-CRF, and achieve the state-of-the-art performance on vessel segmentation in fundus images. Monteiro et al. (2018) implement a 3D version of RNN-CRF for volumetric medical images on top of a V-net segmentation network. The authors evaluate their 3D RNN-CRF on multiple datasets. On the PROMISE 2012 dataset, which consists of MRI prostate images, 3D RNN-CRF improves the Dice from $76.7\% \pm 10.9\%$ to $78.0\% \pm 11.0\%$. And on BraTS 2015, consisting of multimodal MR images of brain tumors, 3D RNN-CRF slightly improves the Dice for tumor segmentation from $73.5\% \pm 10.5\%$ to $73.8\% \pm 10.5\%$. The authors acknowledge that the improvements by 3D RNN-CRF are inconclusive, and attribute that to the intrinsic differences between natural and medical images: 1) object segmentation in 2D RGB images is generally easier with greater contrast and better defined boundaries; 2) the relatively low resolution of 3D volumes causes a mosaic appearance, which poses further challenges on top of blurry edges; and 3) The focal nature of ROIs in medical images downgrades the need to capture global image context beyond what is modeled by the segmentation network, leaving less room for improvement by the CRF. More recently, Chen and de Bruijne (2018) report promising results with their 3D RNN-CRF implementation jointly trained with a 3D U-net for the task of brain lesion segmentation in MR Images. In their approach, CRF operates on high-level features learned by the CNN, which are less prone to image noise than raw intensity values directly taken from the input image. The authors report between six to seven points increase in Dice over a baseline U-Net and other implementations of CRF.

In conclusion, using CRF for post-refinement has proven effective in 2D segmentation tasks, however, its 3D extension shows mixed results. Furthermore, an end-to-end trainable CRF operating on CNN feature maps reduces efforts towards hyper-parameter tuning and is less prone to image noise, hinting at a good direction for future research.

5. Problem II: Weak Annotations

Creating manual segmentation masks, also known as strong annotations, is time consuming and tedious, particularly for 3D images. To combat the high cost associated with strong annotations, researchers have recently explored the use of weak annotations, which can be obtained at significantly lower annotation cost. Reviewing the literature, we have identified three types of weak annotations: 1) image level annotations; 2) sparse annotations, where only a fraction of the slices or pixels are annotated; and 3) model-generated annotations or noisy annotations, which tend to appear under- and over-segmented. Figure 5 shows different types of weak annotations outlined above. While the absence of strong annotations may seem like an obvious handicap, recent research, as summarized below, has shown that it is possible to train fairly effective models with weak annotations.

5.1. Learning with Image Level Labels

Weakly supervised techniques can take advantage of bounding boxes or image level labels. The common weakly supervised approaches suggested for medical image segmentation are based on class activation maps or multiple instance learning. We review both approaches as follows.

5.1.1. Class Activation Maps (CAMs)

A recurring idea in weakly supervised learning is the use of class activation maps Zhou et al. (2016) and its variants (e.g., Selvaraju et al. (2017)), where the idea is to to combine the feature maps to generate class-specific saliency maps. In the following, we review how this technique can be used in conjunction with box-like and image-level annotations.

Image level labels: The trend for tackling the problem of having only image-level labels is to use some form of Class Activation Maps (CAM). Izadyyazdanabadi et al. (2018) use a Multi-Layer Class Activation Map (MLCAM) in the form of a 3-stage inception network where the penultimate feature maps from each network are passed on to the next stage. In parallel CAM followed by global average pooling is applied to these feature maps to obtain the image-level label prediction. The network performance is boosted further by upregulating confident predictions and downregulating uncertain predictions, wherein the regions activated in a single class map

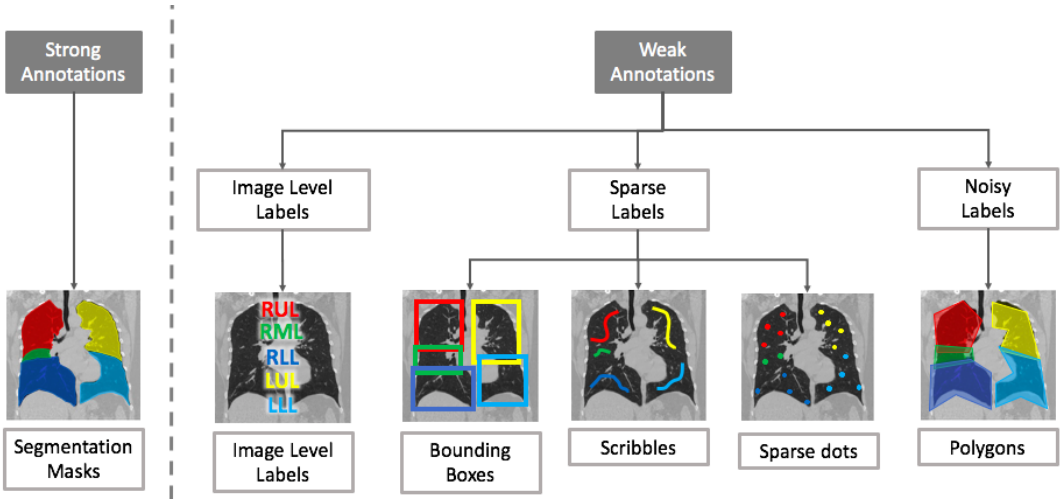


Figure 5: Comparing strong and weak annotations for lung lobes in the coronal view of a chest CT scan. (left) an example of strong annotations in the form of well contoured segmentation masks. (right) examples of different types of weak annotations discussed in this section, which can take the form of (i) bounding box or image level labels, (ii) sparse pixel annotations, or (iii) noisy annotations

are determined to be confident and those activated in both class maps are uncertain. Their average IOU improvement across different tests over the baseline of just using CAM is 20%. Feng et al. (2017) propose a 2-stage approach consisting of a coarse image segmentation followed by a fine instance-level segmentation. The first stage makes use of CAM via an image classification model, which learns whether a slice has a nodule or not. In the second stage, a region of interest is selected around each localized instance in the class activation map and everything outside this region is masked out. Each masked image is then passed to the same classification network to obtain an instance-level segmentation mask, removing false positive regions produced in stage 1. The suggested method achieves a 10% Dice score improvement over the CAM baseline.

5.1.2. Multiple Instance Learning

Multiple Instance Learning (MIL) refers to a classification scheme where the labels are provided for each bag of instances rather than each individual instance. If the label is positive, at least one instance in the bag is positive, but if the label is negative, all instances in the bag are negative. In the context of weakly supervised image segmentation, each image can be considered as a bag of instances where each instance can be a pixel or a tile in the image. By learning to classify the image as a whole, MIL learns to predict instance-level predictions, or equivalently the segmentation prediction for the input image.

Jia et al. (2017) use a multiple instance learning approach to make pixel-level predictions given only image-level cancer labels for histopathology images. The suggested model consists of a VGG network that generates image level classification scores at multiple levels. For this purpose, the authors propose a soft aggregation layer that reduces the feature maps to a cancer classification score. By classifying each image into cancerous and healthy, the authors treat each histopathology image as a bag and each individual pixel in the image as an instance. To regularize training, the authors further impose area constraints (provided during annotation) on segmentation results by MIL. The improvement in F-measure over the baseline (without area constraints and multiple-level predictions) is 6% around the boundaries and 2% overall.

Campanella et al. (2019) train a weakly supervised segmentation model for whole slide histopathology images using only slide-level labels. The slides are divided into small tiles and the multiple instance training procedure includes a full inference pass of the dataset through a CNN, to rank the tiles according to their probability of being positive. The most suspicious tiles in each slide are sequentially passed to an RNN to predict the final slide-level classification. The heatmaps produced by the RNN are considered as segmentation predictions, however, the authors only measure the classification performance since no segmentation masks are available for these datasets. They achieved a 1% increase in AUC when using the RNN classifier over just using the MIL approach directly.

5.2. Learning with Sparse Labels

Incomplete or sparse annotations refer to annotations where the masks are only provided for a fraction of the slices of a 3D volume, or for only a fraction of the pixels of a 2D image. There are various methods for dealing with these annotations as discussed below. Since the pixels (or voxels) are only partially labelled, the underlying theme in these methodologies is the use of a selective pixel loss wherein only the labeled pixels contribute to the loss.

5.2.1. Selective Loss with Mask Completion

The papers discussed in this subsection all attempt to artificially reconstruct the incomplete regions of the ground truth masks and use the completed masks for training.

Zhang et al. (2019c) propose a method for brain extraction and brain tissue segmentation in 3D MR images with a sparse set of annotations, where only a fraction of the slices are annotated at irregular intervals. To complete the sparse annotations,

the authors use active learning. Specifically, the non-annotated slices are ranked by the Dice similarity coefficient between the output feature maps and attention maps, which are generated by a segmentation network equipped with channel-wise and spatial attention mechanisms. The slices with lowest Dice similarity are then presented to an expert for annotation. For the task of brain extraction, the suggested method trained with just 15% of the slices labeled yields results similar to that of a model trained with a 50% annotation rate. For brain tissue segmentation, the model requires 30% of slices to be labeled in order to achieve a similar level of performance. In addition to having spatially sparse annotations, when dealing with time sequence data, the annotations may be temporally sparse. Bai et al. (2018) propose an image sequence segmentation algorithm by combining a fully convolutional network with a recurrent neural network, which incorporates both spatial and temporal information into the segmentation task. The missing annotations are then recovered through a non-rigid label propagation scheme. The model trained using the additional masks recovered for unlabeled time frames achieves 1% improvement in Dice over the U-Net baseline that was only trained on labeled time frames.

Cai et al. (2018) train a 3D segmentation model using only 2D annotations, which consist of diameter markings along the short and long axes of each lesion on the slice where the lesion appears the largest. The authors first refine the initial ground truth markings through the application of GrabCut. The suggested method then alternates between two steps. The first step consists of training the model using augmented ground truth masks. The second step expands the ground truth masks by running the trained model on the non-annotated slices adjacent to the annotated slices followed by applying GrabCut on the generated masks. This iterative process continues until all slices that contain a lesion are annotated. This method achieves 10% improvement in Dice over a model trained using the original diameter markings.

Scribbles have been recognized as a user-friendly alternative to overlapping bounding boxes. To generate the initial ground truth, Can et al. (2018) use a random walk image segmentation approach with a high threshold to perform region growing around seed scribble annotations. Once the initial masks are generated, they suggest an iterative framework to incrementally refine the segmentation masks. Each iteration consists of two stages: 1) training the segmentation network with a CRF-RNN layer using the current annotations, 2) using test time dropout to make multiple predictions for each image and assess the uncertainty of each pixel-level prediction in the image. The new ground truth in each iteration is comprised of only the certain predictions. They achieve a 3% average Dice improvement over their non-iterative baseline.

Unlike previous iterative approaches that attempt to generate the missing regions of the ground truth via Markov Random Fields (MRF) or Conditional Random Fields (CRF) and then use the completed masks for training, Tang et al. (2018) introduce regularization losses that incorporate a CRF in the loss function, thereby eliminating the need for a separate post-processing CRF. The authors also create a kernel cut loss to eliminate the need for a separate step of using graph cuts. These regularization losses are used in addition to the selective segmentation loss which compares the predicted output with the incomplete ground truth. They are able to achieve a 1% improvement in Dice over the other weakly supervised methods that alternate between network training and proposal (training mask) updates.

5.2.2. *Selective Loss without Mask Completion*

Reconstructing the complete segmentation mask is not always a requirement. The papers reviewed below circumvent ground truth completion by modifying the objective function. Silvestri and Antiga (2018) recommend using a hexagonal grid for sparse annotations and show that a dense mask is not a requirement for training an effective pancreas segmentation model for abdominal CT scans. They compare using grids of different strides and the effect of padding the grid points to generate the training masks. The padding process consists of extending the ground truth masks around the sparse points to form discrete label blocks, but does not attempt to complete the segmentation masks. Their results show that using a grid stride of 9 pixels achieve a comparable performance to using a grid size of 3 pixels. The higher the stride, the fewer grid points that need to be annotated. Çiçek et al. (2016) train a 3D model to segment kidney tubules in 3D confocal microscopy images using only sparse annotations. They propose using an additional class for unlabeled pixels. For annotated slices, the segmentation loss is a class-balanced cross entropy where only the labeled pixels contributed to loss. The authors were able to considerably raise their model performance from 0.4 IoU to 0.86 IoU when the annotation rate increases from 2.5% to just 8.9%.

Bokhorst et al. (2018) compare 2 different class-balancing methods that can be used to improve the segmentation performance given sparse annotations without trying to fill in the missing mask pixels. In the suggested method, only the labeled pixels contribute to a weighted segmentation loss. The loss-weighting to balance the classes is performed at the instance level or mini-batch level. The authors show that using instance-based balancing improves the dice score by 1% and mini-batch balancing improves it by 4% when trained entirely on sparsely annotated images.

5.3. *Learning with Noisy Labels*

Noisy labels for the task of image segmentation refer to ambiguities or inaccuracies in the boundaries of the segmentation masks. For medical images in particular, label noise could be induced by annotators unintentionally (random errors), or by inconsistencies between different readers due to human subjectivity concerning ambiguous lesions (expertise errors) Gu et al. (2018). This type of annotation noise can be simulated by representing the labels as a polygon and then reducing the number of polygon vertices, which has the effect of creating segmentation error in the peripheral areas of segmentation masks. Label noise could also arise in the self-learning paradigm (Section 4.4.2) where the model learns

from its own predictions on unlabeled data. Label noise, if left untreated, can degrade the performance of the segmentation model. It is therefore important to utilize strategies that mitigate the adverse effects of label noise during training.

5.3.1. Robust Loss without Mask Refinement

Mirikharaji et al. (2019) propose a learning algorithm resilient to the label noise in the segmentation masks. The suggested method consists of a weighted cross entropy loss function where the contribution of each pixel to the total loss is controlled by model's perception of the annotation quality for the pixels. During training, the weight matrices are updated based on the batches of images with clean annotations, and then used to scale the segmentation loss at the pixel-level for batches with noisy annotations. The authors simulate the annotation noise by replacing the segmentation masks with polygons of varying number of vertices. For the task of skin lesion segmentation, the authors show that a model trained using the suggested loss and 3-vertex polygon masks performs comparably to the model trained using full annotation masks.

5.3.2. Robust Loss with Iterative Mask Refinement

A proper handling of label noise is also studied in the context of self-learning where the model-generated annotations, commonly corrupted by noise, are used to fine-tune the model in an iterative manner. Min et al. (2018) propose a two stream network with independent weights whose concord determine the quality of segmentation mask. Nie et al. (2018) propose a segmentation network with adversarial loss where the job of the discriminator network is to identify the reliable annotated regions from noisy annotations. Readers can refer to Section 4.4.2 for more detailed discussion of these approaches.

6. Discussion

Table 7 presents a summary of the methodologies suggested for the problems of scarce and weak annotations. For clarity, the table is split into two sections, each focusing on one annotation problem. We have grouped the methodologies in each section by the underlying strategy, allowing the readers to find solutions with a similar approach to the problem in one place. We have further used a color encoding to group the methodologies by their required data resources. We hope this table can serve as a strategy guideline, assisting the readers in choosing the right methodology according to the dataset problems they face and the data resources they have available. In what follows, we highlight the important messages of Table 7.

As indicated by the color encoding, the methodologies suggested for the problem of scarce annotations can be placed in three broad categories:

1. *Solutions with low requirements:* This group of solutions rely solely on the available labeled segmentation dataset, requiring no additional labeled or unlabeled training data. Therefore, they should be utilized wherever possible. Of the suggested methodologies, CRF-based post-processing has shown mixed results for 3D segmentation, and altered 3D image representations have achieved medium gains at the price of training several 2D models. Therefore, in addition to traditional data augmentation, which is the de facto solution to the scare annotation problem, we recommend using shape regularization and same-domain data synthesis for both 2D and 3D applications, and CRF-based post-processing for 2D applications.

2. *Solutions with medium requirements:* This set of methodologies requires access to additional labeled or unlabeled training data from the same or a similar domain. Therefore, depending on the application at hand and the availability of the corresponding auxiliary datasets, these solutions may or may not be applicable. Of the suggested methodologies, self-learning approaches have shown mixed results with the exceptions being methods that adopt advanced architectures to handle annotation noise in model-generated annotations. Domain adaptation techniques are effective, but they can be difficult to adopt due to the instability of adversarial training at the core of these methodologies. Semi-supervised learning methods require only additional unlabeled data and are typically less demanding to implement compared to unsupervised domain adaptation methods. Multi-task learning and dataset fusions are both straightforward solutions with reasonable performance gains. In our opinion, self-supervised learning is one of the most promising approaches in this category, requiring only unlabeled data and typically only minor modifications to the architecture.

3. *Solutions with high requirements:* These solutions require access to medical experts, but their elegance lies in the use of expert knowledge in a cost-effective manner. Two solutions in this category are active learning and interactive segmentation where the former determines which samples to be annotated by experts whereas the latter helps experts complete the annotation tasks quickly. If our hands are forced into annotating more data or if additional data annotation is deemed highly advantageous, then these two methodologies should be prioritized in practice.

The methodologies suggested for handling weak annotations are closely related to the types of annotations that are readily available for training. For each type, we compare and recommend methodologies that best suit the given limitation from the perspectives of performance gains and annotation cost.

1. *Noisy annotations:* A common problem with medical datasets and in particular segmentation datasets is annotation noise where the annotated contours may not always follow the contours of the region of interest. Handling annotation noise is important, because not only does it reduce the adverse effects of inter-observer

Table 7: Top-down overview of the methodologies suggested for the problems of scarce and weak annotations, where the methodologies are grouped by the underlying general and specific strategies. We have further used a color encoding to group the methodologies by their required data resources. Methodologies highlighted in green require no further data resources in addition to the original limited annotated dataset available for training; thus, they should be used wherever possible. Methodologies highlighted in orange require access to additional unlabeled data from the same domain or labeled data from a similar domain. Methodologies highlighted in red require experts in the loop; and thus, may not always be a viable option.

Problem I: Scarce Annotations			
General Strategy	Specific Strategy	Methodology	Description
Expanding the dataset	Augmenting the limited data with new artificial examples	Same-domain data synthesis	Training a segmentation model with additional labeled data generated by a data synthesis model
		Traditional data augmentation	Training a segmentation model with additional labeled data generated by spatial and intensity transformation
	Leveraging additional unlabeled data from the same domain	Self learning	Annotating unlabeled images using models' own predictions and then using the augmented dataset for training a segmentation model
		Semi-supervised learning	Training a segmentation model with both labeled and unlabeled data
		Self-supervised learning	Pre-training a model using unlabeled medical data and then fine-tuning the model for the target segmentation task
	Leveraging external labeled data from a similar domain	Dataset fusion	Training a universal segmentation model from heterogeneous datasets by learning to discriminate between the datasets
		Domain adaptation w/ target labels	Training a segmentation model using shared feature representations red across multiple domains
		Domain adaptation w/o target labels	Training a segmentation model using only source domain labels by translating from one domain to the other
	Collecting additional annotations with experts in the loop	Active learning	Selecting unlabeled images for annotation judiciously based on model predictions
		Interactive segmentation	Accelerating the annotation process by propagating the user changes throughout the segmentation mask
Training w/ regularization	Leveraging additional tasks	Multi-task learning	Training a segmentation model with additional heads, each for a separate classification task
	Imposing additional constraints	Shape regularization	Training a segmentation model by imposing shape constraints on predicted segmentation masks
	Leveraging more informative or compressed input data	Altered image representation	Training a segmentation model with a more compact or informative image representation
Post-training refinement	Using post-processing methods to refine segmentations	CRF-based post segmentation	Using CRF as a post-processing or as a trainable module in the segmentation network
Problem II: Weak Annotations			
Leveraging weak annotations	Learning with sparse annotations	Selective loss w/ and w/o mask completion	Training a segmentation model by excluding unannotated pixels from backpropagation
	Learning with noisy annotations	Robust loss w/ and w/o iterative label refinement	Training a segmentation model with mechanisms that downgrade unreliable annotations during training
	Learning with image-level annotations	Class activation maps	Training a classification model with global average pooling and using activation maps as class-specific segmentation
		Multiple instance learning	Training a classification model with aggregation layers and using activation maps as class-specific segmentation

annotation variability on the trained model, but it also enables training with only rough annotations, which can be obtained in a cost-effective manner with significantly shorter annotation time than that of accurate annotations. For instance, the work by Mirikharaji et al. (2019) shows that, with a noise-resilient approach, a skin segmentation model trained with 3-vertex contours can achieve similar performance to a model trained using accurate segmentation masks. Handling annotation noise in medical segmentation datasets is still a fairly new topic and deserves further investigation.

2. *Sparse annotations:* Of the weakly supervised approaches reviewed, the papers tackling sparse annotation have achieved the closest performance to their strongly supervised counterparts; however, the application of sparse annotations may not always be viable. For instance, while dot grids Silvestri and Antiga (2018) may be useful for larger organ segmentation, they would not be as effective for segmenting small lesions. Furthermore, even though sparse annotations are easier to obtain than strong segmentation masks, the annotation process is still not entirely user-friendly, and the training schemes tend to be iterative, leading to longer training periods.

3. *Image- or box-level annotations:* Of the weak annotations reviewed, image-level annotation incur the least annotation cost. Comparing the suggested methodologies, we would recommend using the modified CAM-based approaches with image level-labels. Not only do they use the least expensive form of annotation, but they also show large improvement in Dice over the direct CAM approaches and only fall a couple of Dice points short of using full supervision with strong annotations Feng et al. (2017).

7. Conclusion

This survey presented a detailed review of major data limitations associated with medical image segmentation, namely, scare annotations and weak annotations. For the problem of scarce annotations, we reviewed a diverse set of solutions, ranging from semi-automated solutions that require human experts in the loop such as active learning and interactive segmentation, to fully-automated solutions that leverage unlabeled and synthetic data from the same domain or labeled data

from similar domains. For the problem of weak annotations, we studied solutions with the capability of handling sparse, noisy, or only image-level annotations. We further compared the suggested methodologies in terms of required data resources, difficulty of implementation, and performance gains, highlighting methodologies with the best cost-gain trade-off. We hope this survey increases the community awareness of the strategies for handling the limitations of segmentation datasets and further inspires efforts in this impactful area of research.

8. Acknowledgment

We would like to thank Ju Hu for helping us with compiling the list of related works in the initial stage of this research.

References

Alex, V., Vaidhya, K., Thirunavukkarasu, S., Kesavadas, C., Krishnamurthi, G., 2017. Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation. Journal of Medical Imaging 4, 041311.

Alom, M.Z., Hasan, M., Yakopcic, C., Taha, T.M., Asari, V.K., 2018. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. arXiv preprint arXiv:1802.06955 .

Angermann, C., Haltmeier, M., Steiger, R., Pereverzyev Jr, S., Gizewski, E., 2019. Projection-based 2.5 d u-net architecture for fast volumetric segmentation. arXiv preprint arXiv:1902.00347 .

Bai, W., Chen, C., Tarroni, G., Duan, J., Guitton, F., Petersen, S.E., Guo, Y., Matthews, P.M., Rueckert, D., 2019. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. arXiv preprint arXiv:1907.02757 .

Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D., 2017. Semi-supervised learning for network-based cardiac mr image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 253–260.

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

Baur, C., Albarqouni, S., Navab, N., 2017. Semi-supervised deep learning for fully convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 311–319.

Bokhorst, J.M., Pinckaers, H., van Zwam, P., Nagtegaal, I., van der Laak, J., Ciompi, F., 2018. Learning from sparsely annotated data for semantic segmentation in histopathology images .

Boykov, Y.Y., Jolly, M.P., 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images, in: Proceedings eighth IEEE international conference on computer vision. ICCV 2001, IEEE. pp. 105–112.

Cai, J., Lu, L., Zhang, Z., Xing, F., Yang, L., Yin, Q., 2016. Pancreas segmentation in mri using graph-based decision fusion on convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 442–450.

Cai, J., Tang, Y., Lu, L., Harrison, A.P., Yan, K., Xiao, J., Yang, L., Summers, R.M., 2018. Accurate weakly supervised deep lesion segmentation on ct scans: Self-paced 3d mask generation from recist. arXiv preprint arXiv:1801.08614 .

Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., Fuchs, T.J., 2019. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine URL: https://doi.org/10.1038/s41591-019-0508-1, doi:10.1038/s41591-019-0508-1.

Can, Y.B., Chaitanya, K., Mustafa, B., Koch, L.M., Konukoglu, E., Baumgartner, C.F., 2018. Learning to segment medical images with scribble-supervision alone, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 236–244.

Chartsias, A., Joyce, T., Dharmakumar, R., Tsaftaris, S.A., 2017. Adversarial image synthesis for unpaired multi-modal cardiac data, in: International Workshop on Simulation and Synthesis in Medical Imaging, Springer. pp. 3–13.

Chartsias, A., Joyce, T., Papanastasiou, G., Semple, S., Williams, M., Newby, D., Dharmakumar, R., Tsaftaris, S.A., 2018. Factorised spatial representation learning: application in semi-supervised myocardial segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 490–498.

Chen, C., Dou, Q., Chen, H., Heng, P.A., 2018. Semantic-aware generative adversarial nets for unsupervised domain adaptation in chest x-ray segmentation, in: International Workshop on Machine Learning in Medical Imaging, Springer. pp. 143–151.

Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A., 2019. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. arXiv preprint arXiv:1901.08211 .

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence 40, 834–848.

Chen, S., de Bruijne, M., 2018. An end-to-end approach to semantic segmentation with 3d cnn and posterior-crf in medical images. arXiv preprint arXiv:1811.03549 .

Cheplygina, V., de Bruijne, M., Pluim, J.P., 2019. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Medical Image Analysis .

Christ, P.F., Elshaer, M.E.A., Ettlinger, F., Tatavarty, S., Bickel, M., Bilic, P., Rempfler, M., Armbruster, M., Hofmann, F., D'Anastasi, M., et al., 2016. Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 415–423.

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

Dalca, A.V., Guttag, J., Sabuncu, M.R., 2018. Anatomical priors in convolutional networks for unsupervised biomedical segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9290–9299.

Dmitriev, K., Kaufman, A.E., 2019. Learning multi-class segmentations from single-class datasets, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9501–9511.

Dong, H., Yang, G., Liu, F., Mo, Y., Guo, Y., 2017. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks, in: annual conference on medical image understanding and analysis, Springer. pp. 506–517.

Dou, Q., Ouyang, C., Chen, C., Chen, H., Glocker, B., Zhuang, X., Heng, P.A., 2018. Pnp-adanet: Plug-and-play adversarial domain adaptation network with a benchmark at cross-modality cardiac segmentation. arXiv preprint arXiv:1812.07907 .

Feng, X., Yang, J., Laine, A.F., Angelini, E.D., 2017. Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 568–576.

Fu, C., Ho, D.J., Han, S., Salama, P., Dunn, K.W., Delp, E.J., 2017. Nuclei segmentation of fluorescence microscopy images using convolutional neural networks, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE. pp. 704–708.

Fu, H., Cheng, J., Xu, Y., Wong, D.W.K., Liu, J., Cao, X., 2018. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. IEEE transactions on medical imaging 37, 1597–1605.

Fu, H., Xu, Y., Lin, S., Wong, D.W.K., Liu, J., 2016a. Deepvessel: Retinal vessel segmentation via deep learning and conditional random field, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 132–139.

Fu, H., Xu, Y., Wong, D.W.K., Liu, J., 2016b. Retinal vessel segmentation via deep learning network and fully-connected conditional random fields, in: 2016 IEEE 13th international symposium on biomedical imaging (ISBI), IEEE. pp. 698–701.

Gao, M., Xu, Z., Lu, L., Wu, A., Nogues, I., Summers, R.M., Mollura, D.J.,

2016. Segmentation label propagation using deep convolutional neural networks and dense conditional random field, in: 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), IEEE. pp. 1265–1268.

Ghafoorian, M., Karssemeijer, N., Heskes, T., van Uden, I.W., Sanchez, C.I., Litjens, G., de Leeuw, F.E., van Ginneken, B., Marchiori, E., Platel, B., 2017. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. Scientific Reports 7, 5110.

Giger, M.L., 2018. Whole Brain Segmentation and Labeling from CT Using Synthetic MR Images. Journal of the American College of Radiology 15, 512–520. doi:10.1016/j.jacr.2017.12.028.

Gildenblat, J., Klaiman, E., 2019. Self-supervised similarity learning for digital pathology. arXiv preprint arXiv:1905.08139 .

Gorriz, M., Carlier, A., Faure, E., Giro-i Nieto, X., 2017. Cost-effective active learning for melanoma segmentation. arXiv preprint arXiv:1711.09168 .

Gu, Y., Member, S., Yang, J., Yang, G.z., 2018. Reliable Label-Efficient Learning for Biomedical Image Recognition 9294, 1–11. doi:10.1109/TBME.2018.2889915.

Guibas, J.T., Virdi, T.S., Li, P.S., 2017. Synthetic medical images from dual generative adversarial networks. arXiv preprint arXiv:1709.01872 .

Gupta, S., Hoffman, J., Malik, J., 2016. Cross modal distillation for supervision transfer, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2827–2836.

Harouni, A., Karargyris, A., Negahdar, M., Beymer, D., Syeda-Mahmood, T., 2018a. Universal multi-modal deep network for classification and segmentation of medical images. Proceedings - International Symposium on Biomedical Imaging 2018-April, 872–876. doi:10.1109/ISBI.2018.8363710.

Harouni, A., Karargyris, A., Negahdar, M., Beymer, D., Syeda-Mahmood, T., 2018b. Universal multi-modal deep network for classification and segmentation of medical images, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), pp. 872–876. doi:10.1109/ISBI.2018.8363710.

He, K., Gkioxari, G., Dollr, P., Girshick, R., 2017. Mask R-CNN. arXiv:1703.06870 [cs] URL: http://arxiv.org/abs/1703.06870. arXiv: 1703.06870.

Hesamian, M.H., Jia, W., He, X., Kennedy, P., 2019. Deep learning techniques for medical image segmentation: Achievements and challenges. Journal of digital imaging , 1–15.

Hoo-Chang, S., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D., Summers, R.M., 2016. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35, 1285.

Huang, Y.J., Dou, Q., Wang, Z.X., Liu, L.Z., Jin, Y., Li, C.F., Wang, L., Chen, H., Xu, R.H., 2018. 3d RoI-aware U-Net for Accurate and Efficient Colorectal Tumor Segmentation. arXiv:1806.10342 [cs] URL: http://arxiv.org/abs/1806.10342. arXiv: 1806.10342.

Huo, Y., Xu, Z., Bao, S., Assad, A., Abramson, R.G., Landman, B.A., 2018a. Adversarial synthesis learning enables segmentation without target modality ground truth, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 1217–1220.

Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T.K., Savona, M.R., Abramson, R.G., Landman, B.A., 2018b. Synseg-net: Synthetic segmentation without target modality ground truth. IEEE transactions on medical imaging 38, 1016–1025.

Izadyyazdanabadi, M., Belykh, E., Cavallo, C., Zhao, X., Gandhi, S., Moreira, L.B., Eschbacher, J., Nakaji, P., Preul, M.C., Yang, Y., 2018. Weakly-supervised learning-based feature localization for confocal laser endomicroscopy glioma images, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 300–308.

Jaeger, P.F., Kohl, S.A.A., Bickelhaupt, S., Isensee, F., Kuder, T.A., Schlemmer, H.P., Maier-Hein, K.H., 2018. Retina U-Net: Embarrassingly Simple Exploitation of Segmentation Supervision for Medical Object Detection. arXiv:1811.08661 [cs] URL: http://arxiv.org/abs/1811.08661. arXiv: 1811.08661.

Jamaludin, A., Kadir, T., Zisserman, A., 2017. Self-supervised learning for spinal mris, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 294–302.

Jia, Z., Huang, X., Eric, I., Chang, C., Xu, Y., 2017. Constrained deep weak supervision for histopathology image segmentation. IEEE transactions on medical imaging 36, 2376–2388.

Jin, D., Xu, Z., Tang, Y., Harrison, A.P., Mollura, D.J., 2018. Ct-realistic lung nodule simulation from 3d conditional generative adversarial networks for robust lung segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 732–740.

Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2017. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis 36, 61–78.

Kervadec, H., Bouchtiba, J., Desrosiers, C., Granger, E., Dolz, J., Ayed, I.B., 2019. Boundary loss for highly unbalanced segmentation, in: International Conference on Medical Imaging with Deep Learning, pp. 285–296.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials, in: Advances in neural information processing systems, pp. 109–117.

Kuo, W., Häne, C., Yuh, E., Mukherjee, P., Malik, J., 2018. Cost-sensitive active learning for intracranial hemorrhage detection, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 715–723.

Li, X., Chen, H., Qi, X., Dou, Q., Fu, C.W., Heng, P.A., 2018. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. IEEE transactions on medical imaging 37, 2663–2674.

Li, X., Yu, L., Chen, H., Fu, C.W., Heng, P.A., 2019. Transformation consistent self-ensembling model for semi-supervised medical image segmentation. arXiv preprint arXiv:1903.00348 .

Liskowski, P., Krawiec, K., 2016. Segmenting retinal blood vessels with deep neural networks. IEEE transactions on medical imaging 35, 2369–2380.

Mahapatra, D., Bozorgtabar, B., Thiran, J.P., Reyes, M., 2018. Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 580–588.

Mehta, S., Mercan, E., Bartlett, J., Weave, D., Elmore, J.G., Shapiro, L., 2018. Y-Net: Joint Segmentation and Classification for Diagnosis of Breast Biopsy Images. arXiv:1806.01313 [cs] URL: http://arxiv.org/abs/1806.01313. arXiv: 1806.01313.

Milletari, F., Navab, N., Ahmadi, S.A., 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation, in: 2016 Fourth International Conference on 3D Vision (3DV), IEEE. pp. 565–571.

Min, S., Chen, X., Zha, Z.J., Wu, F., Zhang, Y., 2018. A two-stream mutual attention network for semi-supervised biomedical segmentation with noisy labels. arXiv preprint arXiv:1807.11719 .

Mirikharaji, Z., Hamarneh, G., 2018. Star shape prior in fully convolutional networks for skin lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 737–745.

Mirikharaji, Z., Yan, Y., Hamarneh, G., 2019. Learning to segment skin lesions from noisy annotations. arXiv preprint arXiv:1906.03815 .

Mondal, A.K., Dolz, J., Desrosiers, C., 2018. Few-shot 3d multi-modal medical image segmentation using generative adversarial learning. arXiv preprint arXiv:1810.12241 .

Monteiro, M., Figueiredo, M.A., Oliveira, A.L., 2018. Conditional random fields as recurrent neural networks for 3d medical imaging segmentation. arXiv preprint arXiv:1807.07464 .

Myronenko, A., 2018. 3d MRI brain tumor segmentation using autoencoder regularization. arXiv:1810.11654 [cs, q-bio] URL: http://arxiv.org/abs/1810.11654. arXiv: 1810.11654.

Nie, D., Gao, Y., Wang, L., Shen, D., 2018. Asdnet: Attention based semi-supervised deep networks for medical image segmentation, in: Frangi, A.F., Schnabel, J.A., Davatzikos, C., Alberola-López, C., Fichtinger, G. (Eds.), Medical Image Computing and Computer Assisted Intervention – MICCAI 2018, Springer International Publishing, Cham. pp. 370–378.

Oktay, O., Ferrante, E., Kamnitsas, K., Heinrich, M., Bai, W., Caballero, J., Cook, S.A., De Marvao, A., Dawes, T., O'Regan, D.P., et al., 2017. Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation. IEEE transactions on medical imaging 37, 384–395.

Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., Glocker, B., Rueckert, D., 2018. Attention U-Net: Learning Where to Look for the Pancreas. arXiv:1804.03999 [cs] URL: http://arxiv.org/abs/1804.03999. arXiv: 1804.03999.

Ozdemir, F., Peng, Z., Tanner, C., Fuernstahl, P., Goksel, O., 2018. Active

learning for segmentation by optimizing content information for maximal entropy, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 183–191.

Radosavovic, I., Dollár, P., Girshick, R., Gkioxari, G., He, K., 2018. Data distillation: Towards omni-supervised learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4119–4128.

Ravishankar, H., Venkataramani, R., Thiruvenkadam, S., Sudhakar, P., Vaidya, V., 2017. Learning and incorporating shape models for semantic segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 203–211.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. arXiv:1505.04597 [cs] URL: http://arxiv.org/abs/1505.04597. arXiv: 1505.04597.

Ross, T., Zimmerer, D., Vemuri, A., Isensee, F., Wiesenfarth, M., Bodenstedt, S., Both, F., Kessler, P., Wagner, M., Müller, B., et al., 2018. Exploiting the potential of unlabeled endoscopic video data with self-supervised learning. International journal of computer assisted radiology and surgery , 1–9.

Roth, H.R., Lu, L., Farag, A., Shin, H.C., Liu, J., Turkbey, E.B., Summers, R.M., 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 556–564.

Roy, A.G., Navab, N., Wachinger, C., 2018. Concurrent spatial and channel squeeze & excitationin fully convolutional networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 421–429.

Sakinis, T., Milletari, F., Roth, H., Korfiatis, P., Kostandy, P., Philbrick, K., Akkus, Z., Xu, Z., Xu, D., Erickson, B.J., 2019. Interactive segmentation of medical images through fully convolutional neural networks. arXiv preprint arXiv:1903.08205 .

Schwing, A.G., Urtasun, R., 2015. Fully connected deep structured networks. arXiv preprint arXiv:1503.02351 .

Sedai, S., Mahapatra, D., Hewavitharanage, S., Maetschke, S., Garnavi, R., 2017. Semi-supervised segmentation of optic cup in retinal fundus images using variational autoencoder, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 75–82.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626.

Shin, H.C., Tenenholtz, N.A., Rogers, J.K., Schwarz, C.G., Senjem, M.L., Gunter, J.L., Andriole, K.P., Michalski, M., 2018. Medical image synthesis for data augmentation and anonymization using generative adversarial networks, in: International Workshop on Simulation and Synthesis in Medical Imaging, Springer. pp. 1–11.

Silvestri, G., Antiga, L., 2018. Stereology as weak supervision for medical image segmentation .

Sirinukunwattana, K., Pluim, J.P., Chen, H., Qi, X., Heng, P.A., Guo, Y.B., Wang, L.Y., Matuszewski, B.J., Bruni, E., Sanchez, U., et al., 2017. Gland segmentation in colon histology images: The glas challenge contest. Medical image analysis 35, 489–502.

Sourati, J., Gholipour, A., Dy, J.G., Kurugol, S., Warfield, S.K., 2018. Active deep learning with fisher information for patch-wise semantic segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 83–91.

Sourati, J., Gholipour, A., Dy, J.G., Tomas-Fernandez, X., Kurugol, S., Warfield, S.K., 2019. Intelligent labeling based on fisher information for medical image segmentation using deep learning. IEEE transactions on medical imaging .

Spitzer, H., Kiwitz, K., Amunts, K., Harmeling, S., Dickscheid, T., 2018. Improving cytoarchitectonic segmentation of human brain areas with self-supervised siamese networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 663–671.

Sun, J., Shi, Y., Gao, Y., Wang, L., Zhou, L., Yang, W., Shen, D., 2018a. Interactive medical image segmentation via point-based interaction and sequential patch learning. arXiv preprint arXiv:1804.10481 .

Sun, L., Fan, Z., Huang, Y., Ding, X., Paisley, J., 2018b. Joint CS-MRI Reconstruction and Segmentation with a Unified Deep Network. arXiv:1805.02165 [cs] URL: http://arxiv.org/abs/1805.02165. arXiv: 1805.02165.

Tajbakhsh, N., Hu, Y., Cao, J., Yan, X., Xiao, Y., Lu, Y., Liang, J., Terzopoulos, D., Ding, X., 2019. Surrogate supervision for medical image analysis: Effective deep learning from limited quantities of labeled data. arXiv preprint arXiv:1901.08707 .

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: Full training or fine tuning? IEEE transactions on medical imaging 35, 1299–1312.

Tang, M., Perazzi, F., Djelouah, A., Ben Ayed, I., Schroers, C., Boykov, Y., 2018. On regularized losses for weakly-supervised cnn segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 507–522.

Valindria, V.V., Pawlowski, N., Rajchl, M., Lavdas, I., Aboagye, E.O., Rockall, A.G., Rueckert, D., Glocker, B., 2018. Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI. Proceedings - 2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018 2018-Janua, 547–556. doi:10.1109/WACV.2018.00066.

Wachinger, C., Reuter, M., Klein, T., 2018. Deepnat: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage 170, 434–445.

Wang, G., Li, W., Zuluaga, M.A., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018a. Interactive medical image segmentation using deep learning with image-specific fine tuning. IEEE transactions on medical imaging 37, 1562–1573.

Wang, G., Zuluaga, M.A., Li, W., Pratt, R., Patel, P.A., Aertsen, M., Doel, T., David, A.L., Deprest, J., Ourselin, S., et al., 2018b. Deepigeos: a deep interactive geodesic framework for medical image segmentation. IEEE transactions on pattern analysis and machine intelligence .

Wang, S., Zhou, M., Gevaert, O., Tang, Z., Dong, D., Liu, Z., Tian, J., 2017. A multi-view deep convolutional neural networks for lung nodule segmentation, in: 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE. pp. 1752–1755.

Xia, Y., Xie, L., Liu, F., Zhu, Z., Fishman, E.K., Yuille, A.L., 2018. Bridging the gap between 2d and 3d organ segmentation with volumetric fusion net, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 445–453.

Yang, L., Zhang, Y., Chen, J., Zhang, S., Chen, D.Z., 2017. Suggestive annotation: A deep active learning framework for biomedical image segmentation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 399–407.

Yi, X., Walia, E., Babyn, P., 2018. Generative adversarial network in medical imaging: A review. arXiv preprint arXiv:1809.07294 .

Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O., 2016. Understanding deep learning requires rethinking generalization. arXiv preprint arXiv:1611.03530 .

Zhang, L., Gopalakrishnan, V., Lu, L., Summers, R.M., Moss, J., Yao, J., 2018a. Self-learning to detect and segment cysts in lung ct images without manual annotation, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 1100–1103.

Zhang, L., Wang, X., Yang, D., Sanford, T., Harmon, S., Turkbey, B., Roth, H., Myronenko, A., Xu, D., Xu, Z., 2019a. When unseen domain generalization is unnecessary? rethinking data augmentation. arXiv preprint arXiv:1906.03347 .

Zhang, P., Wang, F., Zheng, Y., 2017a. Self supervised deep representation learning for fine-grained body part recognition, in: 2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017), IEEE. pp. 578–582.

Zhang, P., Zhong, Y., Deng, Y., Tang, X., Li, X., 2019b. A survey on deep learning of small sample in biomedical image analysis. arXiv preprint arXiv:1908.00473 .

Zhang, Y., Miao, S., Mansi, T., Liao, R., 2018b. Task driven generative modeling for unsupervised domain adaptation: Application to x-ray image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 599–607.

Zhang, Y., Yang, L., Chen, J., Fredericksen, M., Hughes, D.P., Chen, D.Z., 2017b. Deep adversarial networks for biomedical image segmentation utilizing unannotated images, in: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (Eds.), Medical Image Computing and Computer Assisted Intervention MICCAI 2017, Springer International Publishing, Cham. pp. 408–416.

Zhang, Y., Yang, Q., 2017. A Survey on Multi-Task Learning. arXiv:1707.08114 [cs] URL: http://arxiv.org/abs/1707.08114. arXiv: 1707.08114.

Zhang, Z., Li, J., Zhong, Z., Jiao, Z., Gao, X., 2019c. A sparse annotation strategy based on attention-guided active learning for 3d medical image seg-

mentation. arXiv preprint arXiv:1906.07367 .

Zhang, Z., Yang, L., Zheng, Y., 2018c. Translating and segmenting multimodal medical volumes with cycle-and shape-consistency generative adversarial network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9242–9251.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J.V., Dalca, A.V., 2019. Data augmentation using learned transforms for one-shot medical image segmentation. arXiv preprint arXiv:1902.09383 .

Zheng, H., Yang, L., Chen, J., Han, J., Zhang, Y., Liang, P., Zhao, Z., Wang, C., Chen, D.Z., 2019. Biomedical image segmentation via representative annotation .

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H., 2015. Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE international conference on computer vision, pp. 1529–1537.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2921–2929.

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

Zhou, Z., Siddiquee, M.M.R., Tajbakhsh, N., Liang, J., 2018b. Unet++: A nested u-net architecture for medical image segmentation, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.

Zhou, Z., Sodha, V., Rahman Siddiquee, M.M., Feng, R., Tajbakhsh, N., Gotway, M., Liang, J., 2019. Models genesis: Generic autodidactic models for 3d medical image analysis, in: Medical Image Computing and Computer-Assisted Intervention (MICCAI). Springer International Publishing. URL: www.tinyurl.com/ModelsGenesisFullVersion.