CUT OUT THE ANNOTATOR, KEEP THE CUTOUT: BETTER SEGMENTATION WITH WEAK SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Constructing large labeled datasets for training segmentation models is an expensive and labor-intensive process. This is a common challenge in machine learning, adressed by methods that require few or no labeled data points such as few-shot learning (FSL) and weakly-supervised learning (WS). Such techniques, however, have limitations when applied to image segmentation—it is difficult to inject multiple forms of knowledge into FSL and other limited label learning techniques, while WS models struggle to fully exploit rich image information. We propose a framework that fuses limited label learning and weak supervision for segmentation tasks, enabling users to train high-performing segmentation CNNs with very few hand-labeled training points. We use CNNs trained with limited labels as weak sources, requiring a very small set of reference labeled images, and introduce a new WS model that focuses on key areas—areas with contention among noisy labels—of the image to fuse these weak sources. Empirically, we evaluate our proposed approach over seven well-motivated segmentation tasks. We show that our methods can achieve within 2 Dice points compared to fully supervised networks while only requiring five hand-labeled training points. Compared to existing limited label supervision methods, including solo few-shot and data augmentation approaches, our approach improves performance by a mean 5 Dice points of the next best method. Finally, we explore the tradeoffs of these various supervision methods, including how well each approach generalizes to new test sets and how well each approach leverages additional unlabeled training data.

1 INTRODUCTION

Automated image segmentation has seen rapid improvements with recent developments in deep learning (Li et al., 2018; Chen et al., 2017; Milletari et al., 2016). Convolutional neural networks (CNNs) produce high segmentation performance—but require large training datasets. Acquiring these labels is laborious and slow. Hand-segmenting images is particularly difficult for medical images, as segmentation is often 3- or 4-dimensional and requires the expertise and time of radiologists. While large datasets and pre-trained networks exist for natural images, medical image segmentation is a far more targeted task, requiring new training sets for every imaging modality, scanner type, anatomical structure, and patient population. Such difficulties abound, but the significant impact of improved medical image segmentation motivates tackling these challenges (Hesamian et al., 2019).

Many approaches have been proposed to mitigate these difficulties by training networks with fewer labeled examples, which we call limited label learning (LLL). For example, few-shot learning (FSL) techniques use a few labeled examples from a previously unseen class to learn a representation that can label new examples of that class (Shaban et al., 2017; Rakelly et al., 2018). Data augmentation attempts to reduce the needed amount of labeled data by introducing variation into training sets through operations such as affine transforms or learned transformations (e.g. deformations learned by GANs) (Zhao et al., 2019; Eaton-Rosen et al., 2018; Çiçek et al., 2016). However, these LLL approaches have limitations: FSL often relies on the segmentations of other classes being present in training datasets, which is unrealistic for many applications. Effective data augmentation often requires model- and task-specific transformations that are difficult to generalize to all tasks. More generally, bringing all of the known domain expertise or structural information to bear is prohibitive.

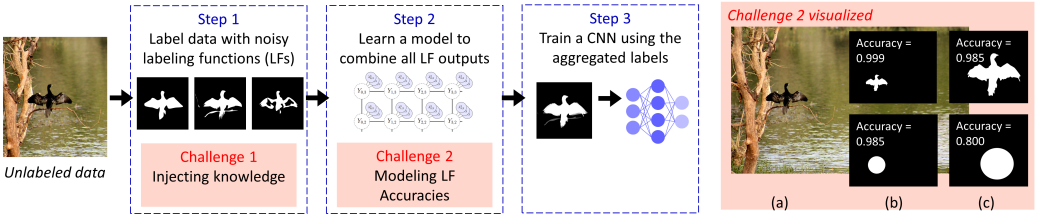

Figure 1: **Left**: Proposed system pipeline (blue) and challenges (red). Multiple LFs noisily label unlabeled data; a label model (LM) then aggregates the LF labels into weak labels that can be used to train a CNN. Two key challenges: how to build LFs that inject knowledge and how to correctly model LF accuracies. **Right**: Second challenge—issues using global accuracy as a metric in the LM: the accuracies of segmentation masks that are clearly correct (b, top) are similar to those that are clearly incorrect (b, bottom). Additionally, segmentation accuracies can be arbitrarily changed by cropping (c).

Weak supervision (WS) is a different approach to training without labels that takes advantage of multiple weak labeling sources to noisily label large training sets (Mintz et al., 2009; Craven & Kumlien, 1999; Takamatsu et al., 2012; Gupta & Manning, 2014; Ratner et al., 2019). The driving idea behind WS is that noisy labels can be generated programatically via labeling functions (LFs), quickly constructing training sets without annotation. These noisy labels are aggregated using a probabilistic graphical model (PGM) to create weak labels for training datasets. Due to the widespread success of weak supervision methods across domains (Dunnmon et al., 2020; Sheng et al., 2020; Bach et al., 2019), weak supervision is an appealing alternative approach for generating segmentation training sets without relying on hand-labeled data. However, existing WS frameworks do not readily extend to the segmentation setting for the following two reasons:

- **Injecting knowledge.** It is unclear how to build LFs that inject knowledge for the particular segmentation task at hand—and that are suitable for any image.
- **Measuring accuracies.** We need to fuse the LFs, but the PGMs of existing WS frameworks rely on estimating the accuracies for each labeling function, which is a problematic metric for segmentation: it is not sensitive to small differences at segmentation borders nor robust to arbitrary image crops (Figure 1, right).

We propose a best-of-both-worlds approach that fuses together LLL techniques and WS for segmentation tasks. The key to our approach is to use LLL techniques as labeling functions. Since WS permits multiple labeling sources, we can bring all available structural information to bear. Further, by relying on LLL models as sources, we obviate the need to craft expertise-injecting labeling functions by hand and exploit the full richness of images. Concretely, we use CNNs trained on a single reference image as our LFs. We resolve the difficulty in tracking accuracies per LF, which are too coarse of a metric for the segmentation setting, by proposing a new WS PGM. Our approach introduces a conditional model that "focuses" on areas of contention, where LFs disagree, more finely tracking the LF performance in such areas.

We validate our approach on seven medical image segmentation tasks. We show that our method can achieve within 2 Dice points of the fully supervised networks using only five labeled data points. We also compare our proposed method to other limited label segmentation approaches, including FSL and data augmentation approaches, and find that our performance exceeds the next best method by a mean 5 Dice points. Finally, we analyze the tradeoffs of our approach and other limited label segmentation approaches. Specifically, we assess how well different approaches generalize to new test sets, and we analyze how various approaches scale with additional unlabeled training data.

2 Related Work

Coarse labels Coarser forms of annotation per image can be used to train segmentation models with less manual labor. For example, bounding boxes, scribbles, and class labels can be used to provide supervision for segmentation models (Sakinis et al., 2019; Xu et al., 2015; Khoreva et al., 2017).

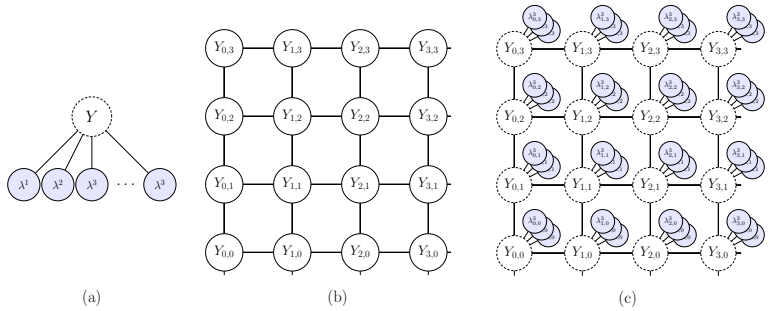

Figure 2: Example latent variable models. (a) A simple PGM with one latent variable Y and several observed labeling functions λ^i. (b) Conventional segmentation grid. (c) The PGM used in this work, where all nodes in a grid of latent variables are weakly labeled by labeling functions.

This approach is sometimes referred to as "weak supervision" in the segmentation literature. We note that this technique is different than the WS approach presented here, as it still requires each training sample to be annotated. We refer to weak supervision as the process by which multiple noisy segmentation masks are generated and aggregated automatically.

Few-shot learning A popular few-shot learning approach enables prediction on previously unseen classes using very few labeled examples of the new class by leveraging representations learned from previous experience (Snell et al., 2017; Sung et al., 2018). Applications to segmentation include Rakelly et al. (2018); Shaban et al. (2017); Dong & Xing (2018). For example, PANet is a proposed approach for few-shot semantic segmentation that utilizes compact representations for each semantic class from the support set, then labels query images by propagating labels from the support set in embedding space (Wang et al., 2019). Few-shot segmentation was extended specifically to the medical imaging case by Roy et al. (2020), where "squeeze and excite" blocks are used to learn class representations without pretrained networks. However, these approaches rely on the availability of labels of other classes during training.

Data augmentation has been shown to enable training on few labeled examples, particularly in medical imaging. These augmentations range from simple operations such as affine transforms, elastic transforms, and contrast jitter (Çiçek et al., 2016; Cireşan et al., 2011) to more complex learned operations, such as the use of GANs to learn intensity and spatial deformation transforms to generate synthetic medical images from a few labeled images (Chaitanya et al., 2020b).

In the appendix, we also discuss semi-supervised learning and conditional random fields, two other popular approaches to segmentation.

3 Segmentation with LLL and WS Fusion

We start with the problem setting and discuss two challenges to using weak supervision in segmentation. We address these with a study of labeling functions and an improved model that fuses limited label learning with weak supervision. We conclude with the learning and inference procedures.

3.1 Problem Setting & Challenges

We are given a training set of n images, $\{\mathbf{X}\}_{i=1}^{i=n}$, where $X_i \in \mathbb{R}^{I,J,K,T}$ is a tensor of P pixels. For each image X_i there is a segmentation mask $Y_i \in \{\mathbf{c}\}^{I,J,K,T}$, where $\{\mathbf{c}\}$ is the set of class labels, yielding the set $\{\mathbf{Y}\}_{i=1}^{i=n}$ of all *unobserved* ground truth segmentation masks. We have access to m labeling functions that provide noisy segmentation masks $\lambda_i^l \in \{\mathbf{c}\}^{I,J,K,T}$ for each X_i. From these observed sources, we seek to estimate the $\{\mathbf{Y}\}_{i=1}^{i=n}$. These provide labels for a large dataset, and we can use this to train any model (e.g., a CNN).

We tackle the binary segmentation case, $\{\mathbf{c}\} = \{-1, 1\}$; simple multiclass extensions are possible.

Model Structure The (unobserved) ground truth mask Y and the (observed) sources λ are jointly modeled by an undirected graphical model (Koller & Friedman, 2009; Wainwright & Jordan, 2008) with graph $G = (V, E)$. Since Y is unobserved, this is a latent-variable model. We use conventional structures for each component and combine them in a natural way:

- In a standard weak supervision approach, the m weak sources are connected to the ground-truth label Y (Fig. 2a), but are conditionally independent on this label,
- The segmentation mask $\{\mathbf{Y}\}$ is a multi-dimensional grid over the pixels (Fig. 2b),
- The full model has one copy of the WS graph for each pixel in the grid (Fig. 2c).

Below, we discuss challenges and our proposed solution for defining a model distribution.

Challenges Our goal is to construct effective LFs and to complete the model specification above by defining the potentials associated with the edges in E and the corresponding parameters. We face two key challenges:

1. **Injecting knowledge via LFs.** The bulk of LFs developed in the literature are for text, and constructing LFs that operate over images is less intuitive.
2. **Refining accuracies via potentials and parameters.** It is not clear how to select potentials and tie parameters in the density function over the graph G when modeling images. On one extreme, separate parameters for each edge in the model in Figure 2c easily results in millions of parameters to be learned. At the other extreme, treating each pixel as independent of all other pixels (as in separate copies of Fig. 2a) results in the problems summarized in Fig. 1 (right): the parameters tend to be equivalent and can be changed by arbitrary crops.

The next two sections tackle these challenges. We start with LF construction.

3.2 Injecting knowledge with LLL-based labeling functions

LF development typically involves writing short programs expressing heuristics (Ratner et al., 2018) or external knowledge base lookups (Mintz et al., 2009; Takamatsu et al., 2012). For segmentation, we observed that while segmentation heuristics are difficult to write, there are instead a host of generic labeling functions we can pull from the CV literature to exploit the structure in segmentation masks and provide weak segmentation signal. We first describe a variety of these simple LFs useful for particular segmentation tasks, then discuss the unified LLL LFs used in this work.

- **Physical models.** For segmentation problems where structures in all images are arranged similarly, physical models (e.g., atlases) can be registered to the unlabeled images and provide weak labels via label propagation.
- **Pre-trained networks.** If the new segmentation task is similar to commonly-studied tasks (e.g. in natural image segmentation), off-the-shelf networks may be deployed.
- **Image features.** For simpler tasks, latent image features (e.g. color, intensity) can provide weak segmentation signal. These features can also be utilized via traditional CV segmentation approaches such as active contour, random walker, and watershed segementation.
- **Clustering.** Unsupervised segmentation methods such as SLIC, which clusters images in XYZ-color space, can be easily implemented for tasks that have distinct boundaries.

LLL labeling functions A unified way to obtain signal or inject knowledge encoded in computer vision models is to apply few-shot or no-shot learning techniques and use the resulting model as a labeling function for weak supervision. Here, we use one-shot learning: we train a CNN from a single labeled data point for each LF. We chose this method of weak labeling because of the variety of segmentation tasks we study; we found that CNNs were adequately flexible as LFs for a variety of different tasks, apt at mimicking the labeling conventions of radiologists, and able to provide strong enough signal to the label model. While at first glance, training multiple neural networks for labeling training data may seem compute-intensive, each training set is only a single image so epochs take only a few seconds to complete. Depending on task difficulty, more sophisticated LLL models could be delivered as an LF.

Our LLL-based LFs depart from the domain-expert LF development approach, which is often repeated for each task. Instead, the same process is followed for all tasks explored here: five images are annotated by hand and used to seed LFs that weakly label the rest of the training set. However,

we note that this approach is probably most suited to structured segmentation problems, such as medical imaging, where there is rich and consistent structure in the images and the labels.

Next, we proceed to the second challenge: selecting our model potentials and tying parameters.

3.3 CONTENTION-BASED PROBABILISTIC MODEL

For our model, we need to define the graph structure G (given in Fig. 2c) and the potentials/parameters associated with the edges in G. The naive approach is to simply model each LF's accuracy parameter—tying these across all of the pixels—but this does not address Challenge 2. However, a separate accuracy per LF per pixel is prohibitively complex. Our approach hits a sweet spot. It exploits a notion of *LF contention*, trading off between the naive model and the (highly-complex) model that has no parameter-tying.

Model Potentials, Parameters, and Contention The latent variable weak supervision model in Fig. 2a expresses the relationship between the sources and the unobserved ground truth label. The potentials are typically those from a binary Ising model (Bach et al., 2017; Ratner et al., 2019; Varma et al., 2019). The density function is defined as,

$$p(Y, \lambda_1, ..., \lambda_m) = \frac{1}{Z} \exp\Big(\theta_Y Y + \sum_i \theta_i \lambda_i Y\Big). \tag{1}$$

Intuitively, the θ_i parameters control the accuracies of the sources. If $\theta_i = 0$, then λ_i is independent of Y—and thus can only output a random guess of Y. As θ_i increases, λ_i and Y vote together more often, and so λ_i is more accurate. For this reason, it is crucial to estimate the parameters of this model; naively combining the noisy labels will over-weight the inaccurate sources.

In our segmentation structure (Fig. 2c), there are many copies of this model. Using unique parameters for each one yields an intractably complex model. Tying parameters naively (i.e., each LF has the same accuracy across all pixels) is too coarse—even a poor LF will typically be very accurate across the entire image. To tackle this, we introduce a notion of LF contention into our density function. Specifically, we use indicator variables that depend on LFs agreeing or disagreeing on a given pixel. Intuitively, this will allow the model to learn a certain set of parameters when the LFs are all agreeing (e.g. image background) vs. when the LFs have some disagreement (e.g. segmentation border). By tying parameters based on LF agreements, this distribution allots additional model parameters to areas of contention without introducing too many free parameters to model the grid-structured PGM. Moreover, this approach enables differential model parameters to be learned even when global LF accuracies are highly similar, and is robust to arbitrary image crops.

We refer to a set of LF agreement conditions over some LF subset D as C_D. C_D exhaustively enumerates each possible combination of LF agreements and disagreements. C_i. For example, the C_D for $D = \{\lambda_1, \lambda_2, \lambda_3\}$ is given by

$$C_D = \{C_1, C_2, C_3, C_4\}$$
$$C_1 = [\lambda_1 \neq \lambda_2, \lambda_1 \neq \lambda_3, \lambda_2 = \lambda_3], C_2 = [\lambda_1 \neq \lambda_2, \lambda_1 = \lambda_3, \lambda_2 \neq \lambda_3]$$
$$C_3 = [\lambda_1 = \lambda_2, \lambda_1 \neq \lambda_3, \lambda_2 \neq \lambda_3], C_4 = [\lambda_1 = \lambda_2, \lambda_1 = \lambda_3, \lambda_2 = \lambda_3]$$

A technical detail we exploit for learning: we always leave two LFs out of the contention set D to ensure sufficient conditional independence for parameter recovery, as we describe next. The set of these left-out LFs is I; if there are $m = 5$ LFs, then $I = \{4, 5\}$.

Formally, we define the density of the contention-based model as

$$p(Y_i, \lambda_i^1, ..., \lambda_i^m) = \frac{1}{Z} \exp\Bigg(\underbrace{\sum_{d=1}^{|C|} \mathbb{1}\{C_d\}}_{\text{Contention term}} \underbrace{\Bigg(\sum_{p=1}^{P} \Big(\theta_{Y_d} Y_i^p + \sum_{k \in D \cup I} \theta_{\lambda_d^k} \lambda_i^{p,k} Y_i^p\Big)\Bigg)}_{\text{Weak supervision}} + \underbrace{\sum_{(p,q) \in E} \theta_{YY_d} Y_i^p Y_i^q}_{\text{Segmentation}}\Bigg). \tag{2}$$

In practice we only sum over a single element of D, which is an equivalent but minimal distribution. enabling backwards mapping as described below.

Learning & Inference Given observed values of labeling functions over a set of images, we wish to learn the parameters θ from (2). Critically, this must be done without relying on ground truth segmentation masks—so that only the λ variables, and not Y, are observed.

We use a simple extension of Fu et al. (2020)'s *method-of-moments*-based approach. For the model (1), $\mathbb{E}[\lambda_i \lambda_j] = \mathbb{E}[\lambda_i Y]\mathbb{E}[\lambda_j Y]$. The left-hand side can be estimated (as we observe the LFs); a system with three such equations, for the pairs $(i, j), (i, k), (j, k)$, determines $\mathbb{E}[\lambda_i Y]$ up to sign. These terms are the *mean* parameters of the model, and under mild technical conditions, they determine the canonical parameters θ (Wainwright & Jordan, 2008). Our extension incorporates the contention term into the decomposition. The full details are in the Appendix.

After solving for the label model parameters, our goal is to compute the marginal probability for each pixel Y_i^p given the observed values of the labeling functions for all pixels in our graph. Various methods have been proposed for performing inference over a large and cyclical graphical model. Here, we approximate the exact conditional using a common graph cut algorithm. We learn the parameters and compute the unary potentials at each pixel, treating each pixel as independent at this stage. The max-flow/min-cut algorithm proposed by Boykov & Kolmogorov (2004) is used to efficiently compute the final segmentation mask considering pairwise interaction between all neighboring pixels. However, we note that many approximate inference algorithms could be used as well, including approximate belief propagation algorithms that would return probabilistic labels.

4 EXPERIMENTS

The primary goal of our evaluation is to show that our proposed weak supervision method can achieve high performance across a range of medical image segmentation tasks (Section 4.2). Secondly, we aim to more deeply understand the benefits of the proposed method compared to other limited label supervision methods (Section 4.3).

4.1 EXPERIMENTAL SETUP

Datasets We select public segmentation datasets to evaluate our proposed method. We chose datasets to include varied imaging modalities, body regions, and structures.

- **ACDC.** Dataset of 100 volumetric short-axis cardiac magnetic resonance (CMR) cine images available for download, totaling 951 2D axial slices. The right ventricle, the left ventricle endocardium, and the myocardium segmentations were provided (Bernard et al., 2018).
- **Abdominal CT.** Dataset of 43 contrast-enhanced abdominal computed tomography (ACT) scans, totalling 10,235 2D axial slices. We use the segmentations provided for the spleen, esophagus, and liver (Gibson et al., 2018; Roth et al., 2016; 2015; Clark et al., 2013).
- **BRATS.** Dataset of 210 head magnetic resonance (HMR) scans, totalling 31,775 axial slices. We use the flair HMR sequence and segment any region positive for tumor (Menze et al., 2014; Bakas et al., 2017; 2018).

Baselines We compare our proposed approach to other limited-label segmentation methods. We select the baselines that have similar inputs (a few labeled images and many unlabeled images) and outputs (a trained segmentation neural network) to our proposed approach, but use a distinct method, and have published implementations by the authors. In each case, we use the same unlabeled and five labeled images we use in our proposed approach to ensure the training datasets are comparable.

- **Few-shot approaches.** We compare against PANet, a leading few-shot learning approach to semantic segmentation developed for natural images. We also compare against Squeeze and Excite ("S&E"), a few-shot approach for medical images.
- **Data augmentation approaches.** We compare against a more sophisticated method of data augmentation that trains generative adversarial networks (GANs) using unlabeled images then transforms the labeled images using the trained GANs to generate a larger, labeled training set ("GAN DA").
- **Traditional supervision.** We train a model using only the five labeled reference images as the lower bound of expected performance ("FS-5"). We also train a fully supervised

Table 1: Performance of proposed method and baselines.

Method	Cardiac MR			Ab. CT			BRATS
	Endo.	Myo.	RV	Liver	Spleen	Esophagus	Tumor
FS-5	0.79	0.56	0.52	0.92	**0.93**	0.48	0.77
PANet	0.62	0.33	0.42	0.58	0.23	0.26	N/A
S&E	0.47	0.21	0.26	0.24	0.05	0.01	N/A
GAN DA	0.91	0.76	0.77	0.87	0.84	0.55	0.74
Proposed	**0.94**	**0.79**	**0.86**	**0.93**	0.91	**0.61**	0.77
FS-all	0.96	0.87	0.94	0.96	0.94	0.76	0.84

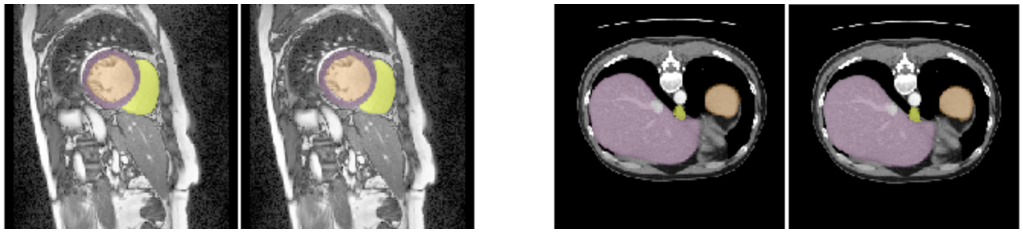

Figure 3: Example results. The left of each pair is the ground truth, the right of each pair is produced by the proposed method. Left: orange is endocardium, pink is myocardium, yellow is right ventricle. Right: orange is spleen, pink is liver, yellow is esophagus.

network using all ground truth segmentations of all training data as the upper bound of expected performance ("FS-all").

Training and Evaluation Each dataset is split into 60% training, 20% validation, and 20% testing. Image preprocessing is described in the Appendix. We train a U-Net (Çiçek et al., 2016) to minimize the binary cross entropy loss using the Adam optimizer with learning rate set to 1e-3. Batch size was set to 16. Random affine transforms, color jitter, and elastic transforms were used during training for all models except FS-5. The end model was trained on axial slices.

The Dice coefficient, $Dice(Y_i, Y_j) = \frac{2|Y_i \cap Y_j|}{|Y_i| + |Y_j|}$, is used to evaluate networks. Each network was trained three times with different random seeds. The mean Dice scores are reported.

4.2 Performance of proposed method

The Dice scores of our proposed approach and all baselines are shown for the seven segmentation tasks in Table 1; visualizations of the segmentations are shown in Figure 3 and in the Appendix in Figure A.1. Despite using only five labeled images, we obtain near fully supervised performance on the endocardium, liver, and spleen segmentations, achieving on average within 0.027 points of the fully supervised performance. On the more difficult segmentation tasks, we still see boost compared to existing limited-label segmentation approaches, though there is larger gap between FS and WS performance. This is to be expected, since the segmentation of smaller structures is more difficult to accomplish. Still, we see strong agreement in the shape and location of the WS segmentations compared to the ground truth segmentations.

4.3 Tradeoffs with other limited label approaches

In addition to the performance comparisons above, we wanted to understand more broadly the tradeoffs associated with different limited label approaches. Specifically, we considered two comparisons: how do different approaches generalize to new test sets, and how do different approaches scale with additional unlabeled data.

Generalization performance For methods that use fewer training labels, we hypothesized that the final CNNs may be less robust to natural image variations than fully supervised networks. We

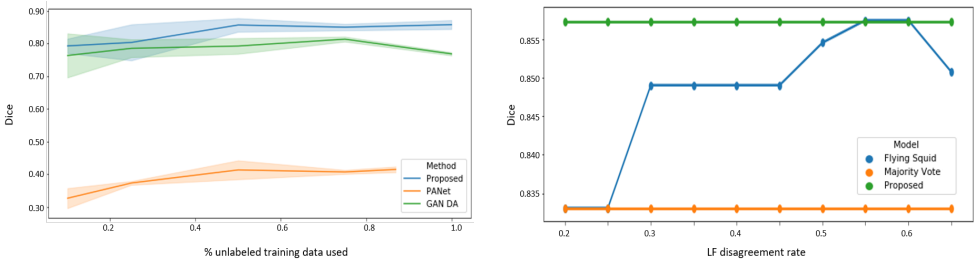

Figure 4: Comparison of models for different: left, amount of unlabeled data; right, image crops.

assessed the models' abilities to generalize to new test sets for two tasks. While the new datasets are of the same structure, view, and modality as the originals, they represent different patient populations and scanners. As an additional CMR test set, we collected 100 CMR images with endocardium segmentations through collaboration with the NIH. As an additional ACT test set, we use images and liver segmentations from the CHAOS challenge (Kavur et al., 2020a;b; 2019).

Table 2: Generalization performance.

	Endo.	Liver
PANet	+3.8%	+1.1%
GAN DA	-15.3%	-5.2%
Proposed	-6.0%	+1.0%
FS-all	-3.9%	-0.00%

We compare the performance of the trained models on these new test sets and report the change in performance compared to the original test set in Table 2. We compared the best performing method from the few-shot approaches (PANet) along with our proposed approach, the data augmentation approach, and the fully supervised network. We see that the WS approach behaves most similarly to the fully supervised networks. However, PANet actually is able to best generalize to the new test sets (relative to its original performance), possibly because of its inference procedure, which utilizes labeled images from the new test set.

Performance scaling with amount of unlabeled data Since unlabeled data is much easier to collect than manually-annotated data, LLL approaches that are able to effectively utilize unlabeled data may have an advantage when deployed to real-world applications. In Figure 4, we show the performance of our proposed approach, the highest performing FSL approach (PANet), and the data augmentation network as more unlabeled data is provided during training for the CMR right ventricle task. We observe that the performance separation between WS and other approaches is larger with more unlabeled data. In the low data regime, WS and GAN-DA perform similarly.

Label model performance To show the advantages of the proposed segmentation LM versus existing WS models, we conduct an experiment with synthetic datasets. We generate a dataset according to Equation 2, and pad this dataset with negatively labeled examples. This process mimics cropping an image around the ROI. We show the performance of our proposed LM, majority vote (MV), and a leading existing WS framework (Flying Squid, Fu et al. (2020)) in Figure 4. We see our proposed model is unaffected by background cropping (or equivalently, LF disagreement rate) while the existing WS tool has varying performance as a function of arbitrary crops. We report the Dice score of MV, the proposed LM, and Flying Squid in the Appendix Table 1.

5 Conclusion

We present a general approach to training segmentation network with limited hand labels. This approach is flexible, capable of utilizing various forms of noisy signal to weakly label training data, which can in turn be used to train any downstream model.

ACKNOWLEDGMENTS

Anon.

REFERENCES

Stephen H. Bach, Bryan He, Alexander J. Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, Sydney, Australia, 2017.

Stephen H Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alex Ratner, Braden Hancock, Houman Alborzi, et al. Snorkel drybell: A case study in deploying weak supervision at industrial scale. In *Proceedings of the 2019 International Conference on Management of Data*, pp. 362–375, 2019.

Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. Self-supervised learning for cardiac mr image segmentation by anatomical position prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 541–549. Springer, 2019.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4: 170117, 2017.

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

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018.

Yuri Boykov and Vladimir Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE transactions on pattern analysis and machine intelligence*, 26(9):1124–1137, 2004.

Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2001.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *arXiv preprint arXiv:2006.10511*, 2020a.

Krishna Chaitanya, Neerav Karani, Christian F Baumgartner, Anton Becker, Olivio Donati, and Ender Konukoglu. Semi-supervised task-driven data augmentation for medical image segmentation. *arXiv preprint arXiv:2007.05363*, 2020b.

Liang Chen, Paul Bentley, Kensaku Mori, Kazunari Misawa, Michitaka Fujiwara, and Daniel Rueckert. Self-supervised learning for medical image analysis using image context restoration. *Medical image analysis*, 58:101539, 2019.

Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020.

Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pp. 424–432. Springer, 2016.

Dan C Cireşan, Ueli Meier, Jonathan Masci, Luca M Gambardella, and Jürgen Schmidhuber. High-performance neural networks for visual object classification. *arXiv preprint arXiv:1102.0183*, 2011.

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al. The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6): 1045–1057, 2013.

M. Craven and J. Kumlien. Constructing biological knowledge bases by extracting information from text sources. In *International Conference on Intelligent Systems for Molecular Biology (ISMB)*, 1999.

Nanqing Dong and Eric P Xing. Few-shot semantic segmentation with prototype learning. In *BMVC*, volume 3, 2018.

Jared A Dunnmon, Alexander J Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew P Lungren, Daniel L Rubin, et al. Cross-modal data programming enables rapid medical machine learning. *Patterns*, pp. 100019, 2020.

Zach Eaton-Rosen, Felix Bragman, Sebastien Ourselin, and M Jorge Cardoso. Improving data augmentation for medical image segmentation. 2018.

Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. *arXiv preprint arXiv:2002.11955*, 2020.

Eli Gibson, Francesco Giganti, Yipeng Hu, Ester Bonmati, Steve Bandula, Kurinchi Gurusamy, Brian Davidson, Stephen P Pereira, Matthew J Clarkson, and Dean C Barratt. Automatic multi-organ segmentation on abdominal ct with dense v-networks. *IEEE transactions on medical imaging*, 37(8):1822–1834, 2018.

Sonal Gupta and Christopher D Manning. Improved pattern learning for bootstrapped entity extraction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pp. 98–108, 2014.

Mohammad Hesam Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of digital imaging*, 32(4):582–596, 2019.

A. Emre Kavur, N. Sinem Gezer, Mustafa Barış, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, Bora Baydar, Dmitry Lachinov, Shuo Han, Josef Pauli, Fabian Isensee, Matthias Perkonigg, Rachana Sathish, Ronnie Rajan, Sinem Aslan, Debdoot Sheet, Gurbandurdy Dovletov, Oliver Speck, Andreas Nürnberger, Klaus H. Maier-Hein, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. CHAOS Challenge - Combined (CT-MR) Healthy Abdominal Organ Segmentation, January 2020a. URL https://arxiv.org/abs/2001.06535.

A. Emre Kavur, Naciye Sinem Gezer, Mustafa Barış, Yusuf Şahin, Savaş Özkan, Bora Baydar, Ulaş Yüksel, Çağlar Kılıkçıer, Şahin Olut, Gözde Bozdağı Akar, Gözde Ünal, Oğuz Dicle, and M. Alper Selver. Comparison of semi-automatic and deep learning based automatic methods for liver segmentation in living liver transplant donors. *Diagnostic and Interventional Radiology*, 26: 11–21, January 2020b. doi: 10.5152/dir.2019.19. URL https://doi.org/10.5152/dir.2019.19025.

Ali Emre Kavur, M. Alper Selver, Oğuz Dicle, Mustafa Barış, and N. Sinem Gezer. CHAOS - Combined (CT-MR) Healthy Abdominal Organ Segmentation Challenge Data, April 2019. URL https://doi.org/10.5281/zenodo.3362844.

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

Pushmeet Kohli, M Pawan Kumar, and Philip HS Torr. P3 & beyond: Solving energies with higher order cliques. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8. IEEE, 2007.

Pushmeet Kohli, Philip HS Torr, et al. Robust higher order potentials for enforcing label consistency. *International Journal of Computer Vision*, 82(3):302–324, 2009.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pp. 109–117, 2011.

Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE transactions on medical imaging*, 37(12):2663–2674, 2018.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571. IEEE, 2016.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pp. 1003–1011. Association for Computational Linguistics, 2009.

Cheng Ouyang, Carlo Biffi, Chen Chen, Turkay Kart, Huaqi Qiu, and Daniel Rueckert. Self-supervision with superpixels: Training few-shot medical image segmentation without annotation. *arXiv preprint arXiv:2007.09886*, 2020.

Kate Rakelly, Evan Shelhamer, Trevor Darrell, Alexei A Efros, and Sergey Levine. Few-shot segmentation propagation with guided networks. *arXiv preprint arXiv:1806.07373*, 2018.

A. J. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, 2019.

A.J. Ratner, S.H. Bach, H. Ehrenberg, J. Fries, S. Wu, and C. Ré. Snorkel: Rapid training data creation with weak supervision. In *VLDB*, 2018.

Holger Roth, Amal Farag, Evrim Turkbey, Le Lu, Jiamin Liu, and Ronald Summers. Data from pancreas-ct. 2016.

Holger R Roth, Le Lu, Amal Farag, Hoo-Chang Shin, Jiamin Liu, Evrim B Turkbey, and Ronald M Summers. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In *International conference on medical image computing and computer-assisted intervention*, pp. 556–564. Springer, 2015.

Abhijit Guha Roy, Shayan Siddiqui, Sebastian Pölsterl, Nassir Navab, and Christian Wachinger. 'squeeze & excite'guided few-shot segmentation of volumetric images. *Medical image analysis*, 59:101587, 2020.

Tomas Sakinis, Fausto Milletari, Holger Roth, Panagiotis Korfiatis, Petro Kostandy, Kenneth Philbrick, Zeynettin Akkus, Ziyue Xu, Daguang Xu, and Bradley J Erickson. Interactive segmentation of medical images through fully convolutional neural networks. *arXiv preprint arXiv:1903.08205*, 2019.

Amirreza Shaban, Shray Bansal, Zhen Liu, Irfan Essa, and Byron Boots. One-shot learning for semantic segmentation. *arXiv preprint arXiv:1709.03410*, 2017.

Ying Sheng, Nguyen Ha Vo, James B Wendt, Sandeep Tata, and Marc Najork. Migrating a privacy-safe information extraction system to a software 2.0 design. 2020.

Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International journal of computer vision*, 81(1):2–23, 2009.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in neural information processing systems*, pp. 4077–4087, 2017.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1199–1208, 2018.

S. Takamatsu, I. Sato, and H. Nakagawa. Reducing wrong labels in distant supervision for relation extraction. In *Meeting of the Association for Computational Linguistics (ACL)*, 2012.

P. Varma, F. Sala, A. He, A. J. Ratner, and C. Ré. Learning dependency structures for weak supervision models. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, 2019.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Kaixin Wang, Jun Hao Liew, Yingtian Zou, Daquan Zhou, and Jiashi Feng. Panet: Few-shot image semantic segmentation with prototype alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9197–9206, 2019.

Jia Xu, Alexander G Schwing, and Raquel Urtasun. Learning to segment under various forms of weak supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3781–3790, 2015.

Amy Zhao, Guha Balakrishnan, Fredo Durand, John V Guttag, and Adrian V Dalca. Data augmentation using learned transformations for one-shot medical image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8543–8553, 2019.

A APPENDIX

We provide additional discussion of related work. Then we give a detailed description of the learning procedure for the model introduced in Section 3.3. Afterwards, we include additional experiments and experimental details.

A.1 EXTENDED RELATED WORK

Semi-supervised learning Self- and semi-supervised methods have been proposed for image segmentation without relying on the availability of other classes during training (Chen et al., 2020; Bai et al., 2019; Chen et al., 2019). Instead, semi-supervised segmentation often pretrains on large unlabeled datasets to learn useful representations without any annotations. For example, contrastive learning has been proposed to pretrain CNNs, which are then fine-tuned using a small number of manually segmented images (Chaitanya et al., 2020a). Superpixel segmentations have been used to train discriminative image representations without labeled data in SSL-ALPNet, which can be used to segment test images without fine tuning (Ouyang et al., 2020).

Conditional random fields Traditional approaches to modeling segmentation masks for images use conditional random fields (CRFs). These distributions often utilize unary potentials over each pixel and pairwise potentials between pixels. Many methods were proposed for defining the unary potentials—including engineering features over the color, location, and texture of regions in the image (Shotton et al., 2009)—as well as pairwise interactions between pixels, such as how smooth segmentations were expected to be (Kohli et al., 2009; Krähenbühl & Koltun, 2011). Additional work was devoted to performing inference over the large CRFs, developing new methods for approximate inference via belief propagation and graph cuts (Kohli et al., 2007; Boykov et al., 2001; Boykov & Kolmogorov, 2004).

Note that the graph structure we use (Fig. 2c) bears resemblance to those used in past work in CRFs. However, in most traditional CRF frameworks, the parameters of the PGM are learned or tuned from training data. Here, we work with a latent ground truth variable Y and generate probabilistic labels for each pixel by learning how to aggregate LF votes without access to ground truth data.

A.2 LEARNING MODEL PARAMETERS

We rely on a technical fact: our proposed distribution is minimal, so that finding the canonical parameters θ can be accomplished by finding the mean parameters of the model: $E[\mathbb{1}\{C_d\}Y_i^p]$, $E[\mathbb{1}\{C_d\}Y_i^pY_i^q]$, and $E[\mathbb{1}\{C_d\}\lambda_i^{p,k}Y_i^p]$ and performing the mean-to-canonical parameter mapping. See Wainwright & Jordan (2008) for more details. We take $E[\mathbb{1}\{C_d\}Y_i^p]$ and $E[\mathbb{1}\{C_d\}Y_i^pY_i^q]$ as inputs, representing the conditional class balance of all pixels and the conditional rate of agreement of neighboring pixels, respectively. We note that these values can be estimated from a validation set, tuned, or estimated from the observed values of the labeling functions. The remaining mean parameters to be estimated are then $E[\mathbb{1}\{C_d\}\lambda_i^{p,k}Y_i^p]$. For conciseness, we will refer to Y_i^p as Y and $\lambda_i^{p,k}$ as λ^k in the remainder of this section.

The mean parameter $E[\mathbb{1}\{C_d\}\lambda^kY]$ can be interpreted as the scaled accuracy of labeling function S^k under condition C_d, a_d^k. That is, rather than being between 0 and 1, it is scaled to be between -1 and 1:

$$a_d^k = E[\mathbb{1}\{C_d\}\lambda^kY] = P((\lambda^k = Y) \cap (C_D)) - P((\lambda^k \neq Y) \cap (C_D)). \tag{3}$$

This mean parameter cannot be directly estimated (i.e., from samples of λ and Y) as we do not observe the ground truth label Y. Instead, building off of the triplet method for parameter recovery presented in Fu et al. (2020), the conditional accuracy of each labeling function can be efficiently computed in closed-form using a system of equations.

We use the following decomposition to factor an observable quantity into the product of two conditional accuracies a_d^j and a_d^k. We take j to be in the contention set C and k to be in the non-contention part I. The decomposition holds due to the fact that these terms are uncorrelated, which follows from

the conditional independence of the underlying λ variables.

$$\begin{aligned} a_d^j a_d^k &= E[\mathbb{1}\{C_d\}\lambda_i^{p,j} Y_i^p] E[\mathbb{1}\{C_d\}\lambda_i^{p,k} Y_i^p] \\ &= E[\mathbb{1}\{C_d\}\lambda_i^{p,j} Y_i^p \mathbb{1}\{C_d\}\lambda_i^{p,k} Y_i^p] \\ &= E[\mathbb{1}\{C_d\}\lambda_i^{p,j}\lambda_i^{p,k}]. \end{aligned}$$

Having access to conditionally independent labeling functions is precisely why we selected two LFs and kept them in the set I, which does not overlap with the contention set D.

With the equation above, we can observe all of the terms such as $\mathbb{1}\{C_d\}\lambda_i^{p,j}\lambda_i^{p,k}$, so that we can empirically estimate the final term. If we form triplets of such equations, for the pairs $(j,k),(j,\ell),(k,\ell)$, where $j \in D$ and $k,\ell \in I$, then we can use the following system of equations to solve for the a_d^j, a_d^k, a_d^ℓ, up to sign. Specifically, the system is

$$\begin{aligned} a_d^j a_d^k &= E[\mathbb{1}\{C_d\}\lambda_i^{p,j}\lambda_i^{p,k}], \\ a_d^j a_d^\ell &= E[\mathbb{1}\{C_d\}\lambda_i^{p,j}\lambda_i^{p,\ell}], \\ a_d^k a_d^\ell &= E[\mathbb{1}\{C_d\}\lambda_i^{p,k}\lambda_i^{p,\ell}]. \end{aligned}$$

The magnitude of the solution is for j

$$|a_d^j| = \sqrt{\frac{E[\mathbb{1}\{C_d\}\lambda_i^{p,j}\lambda_i^{p,k}] \times E[\mathbb{1}\{C_d\}\lambda_i^{p,j}\lambda_i^{p,\ell}]}{E[\mathbb{1}\{C_d\}\lambda_i^{p,k}\lambda_i^{p,\ell}]}}.$$

In practice, we use the empirical versions of each of the expectation terms. In general, as long as the accuracies are better than random, the sign is positive, so that we fully recover a_d^j. Even in cases where the signs of some accuracies are negative, it is still possible to recover in a range of conditions, using the technique detailed in Ratner et al. (2019).

This procedure can be done for any of the LFs and choices of the contention. All that is required is to select valid triplets. By design, there is only one possible valid grouping of triplets: each labeling function in D upon which the conditions rely is grouped with the two labeling functions in I upon which no conditions rely. We take the median conditional accuracy estimate for each labeling function in I. The division of labeling functions into I and D is chosen to maximize the number of LF disagreements caught by the conditionals in the dependent set.

With this we have recovered all of the mean parameters. To obtain the canonical parameters—the θ's in (2)—we use the mean-to-canonical mapping, as detailed in Wainwright & Jordan (2008).

A.3 Additional Experimental Results

Image preprocessing The CMR images were corrected with an N4 bias field correction and normalized with histogram equalization. The ACT images were preprocessed with a CT window (window level = 50, window width = 400) and each axial slice was resized to 224x224, then normalized to the range [0,1]. The BRATS dataset was resized such that each axial slice was 224x224, then normalized to the range [0,1].

Additional Results In Figure A.1, we show qualitative visualizations of the segmentations described in Section 4.2. In Table 1, we give the Dice score of the majority vote of the labeling functions, our proposed label model, and Flying Squid (Fu et al. (2020)) on the validation set. We note that the current implementation of Flying Squid was unable to perform inference over the large number of pixels in the ACT and BRATS datasets.

Table 1: Label model performance.

Method	Cardiac MR			Ab. CT			BRATS
	LV Endo	LV Epi	RV	Liver	Spleen	Kidney	Tumor
Majority vote	0.92	0.77	0.79	0.94	0.91	0.42	0.77
Flying squid	0.92	0.77	0.79	-	-	-	-
Proposed	0.93	0.78	0.82	0.94	0.91	0.45	0.77

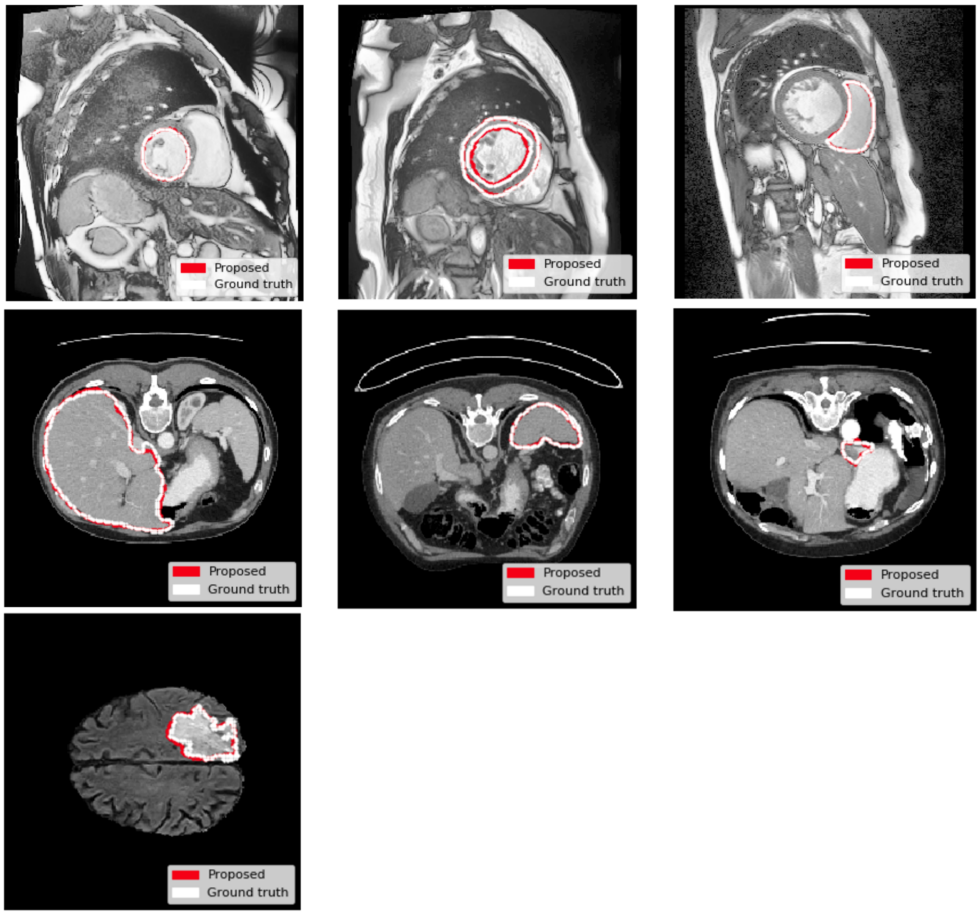

Figure A.1: Example results visualized for each dataset, starting from the top left and proceeding by row: endocardium, myocardium, right ventricle, liver, spleen, esophagus, and tumor.