Size-constraint loss for weakly supervised CNN segmentation

Hoel Kervadec
ÉTS Montréal
hoel.kervadec.1@etsmtl.net

Jose Dolz
ÉTS Montréal

Meng Tang
University of Waterloo, Department of Computer science

Éric Granger
ÉTS Montréal

Yuri Boykov
University of Waterloo, Department of Computer science

Ismail Ben Ayed
ÉTS Montréal

Abstract

Weak supervision, e.g., in the form of partial labels or image tags, is currently attracting significant attention in CNN segmentation as it can mitigate the lack of full and laborious pixel/voxel annotations, a common problem in medical imaging. Embedding high-order (global) inequality constraints on the network output, for instance, on the size of the target region, can leverage unlabeled data, guiding training with domain-specific knowledge. Inequality constraints are very flexible because they do not assume exact prior knowledge. However, constrained Lagrangian optimization has been largely avoided in deep networks, mainly for computational tractability reasons. To the best of our knowledge, the method of Pathak et al. [**?**] is the only prior work that addresses constrained deep CNNs in weakly supervised segmentation. It uses the constraints to synthesize fully-labeled training masks (proposals) from weak labels, mimicking full supervision and facilitating dual optimization.

We propose to introduce a differentiable term, which enforces inequality constraints directly in the loss function, avoiding expensive Lagrangian dual iterates and proposal generation. From constrained-optimization perspective, our simple approach is not optimal as there is no guarantee that the constraints are satisfied. However, surprisingly, it yields *substantially* better results than the proposal-based constrained CNNs in [**?**], while reducing the computational demand for training. In the context of cardiac image segmentation, we reached a segmentation performance close to full supervision while using a fraction of the ground-truth labels (0.1% of the pixels of the ground-truth masks) and image-level tags. Our framework can be

Submitted to 1st Conference on Medical Imaging with Deep Learning (MIDL 2018). Do not distribute.

easily extended to other inequality constraints, e.g., shape moments [?] or region statistics [?]. Therefore, it has the potential to close the gap between weakly and fully supervised learning in semantic medical image segmentation. Our code is publicly available.

1 Introduction

In the recent years, deep convolutional neural networks (CNNs) have been dominating semantic segmentation problems, both in computer vision and medical imaging, achieving ground-breaking performances when full-supervision is available [? ?]. In semantic segmentation, full supervision requires laborious pixel/voxel annotations, which may not be available in a breadth of applications, more so when dealing with volumetric data. Therefore, weak supervision with partial labels, for instance, bounding boxes [?], points [?], scribbles [? ? ?], or image tags [? ?], is attracting significant research attention. Imposing prior knowledge on the the network's output in the form of unsupervised loss terms is a well-established approach in machine learning [? ?]. Such priors can be viewed as regularization terms that leverage unlabeled data, embedding domain-specific knowledge. For instance, our recent studies in [? ?] showed that direct regularization losses, e.g., dense conditional random field (CRF) or pairwise clustering, can yield outstanding results in weakly supervised segmentation, reaching almost full-supervision performances in natural image segmentation; see the results in [?]. Surprisingly, such a principled direct-loss approach is not common in weakly supervised segmentation. In fact, most of the existing techniques synthesize fully-labeled training masks (proposals) from the available partial labels, mimicking full supervision [? ? ? ?]. Typically, such proposal-based techniques iterate two steps: CNN learning and proposal generation facilitated by dense CRFs and fast mean-field inference [?], which are now the de-facto choice for pairwise regularization in semantic segmentation algorithms.

This study continues our line of recent works in [? ?], in which we showed the potential of direct pairwise regularization losses (e.g., dense CRF and Potts) in weakly supervised segmentation. Our purpose here is to embed high-order (global) inequality constraints on the network output directly in the loss function, so as to guide learning. For instance, assume that we have some prior knowledge on the size (volume) of the target region, e.g., in the form of lower and upper bounds on size, a common scenario in medical image segmentation [? ?]. Let $I : \Omega \subset \mathbb{R}^{2,3} \rightarrow \mathbb{R}$ denotes a given training image, and $Y = (Y_1, \ldots, Y_{|\Omega|}) \in \{0,1\}^{|\Omega|}$ its weak (partial) ground-truth segmentation. Y is a weak/partial annotation the target region, e.g., a few points (see the examples in Fig. 2) or image-level tags. In this case, one can optimize a cross-entropy loss subject to inequality constraints on the network outputs:

$$\min_{\boldsymbol{\theta}} \ \mathcal{H}(Y,S) \quad \text{s.t} \quad a \leq \sum_{p \in \Omega} S_p \leq b \tag{1}$$

where S = $(S_1, \ldots, S_{|\Omega|}) \in [0,1]^{|\Omega|}$ is a vector of softmax probabilities generated by the network at each pixel p and $\mathcal{H}(Y,S) = -\sum_{p\in\Omega} Y_p \log(S_p)$. Priors a and b denote the given upper and lower bounds on the size (or cardinality) of the target region. Inequality constraints of the form in (1) are very flexible because they do not assume exact knowledge of the target size, unlike [? ?]. Also, multiple instance learning constraints [?], which enforce image-tag priors, can be handled by constrained model (1). Image tag is a form of weak supervision, which enforces whether a target region is present or absent in a given training image [?].

Even though constraints of the form (1) are linear (and hence convex) with respect to the network outputs, constrained problem (1) is very challenging due to the non-convexity of CNNs. One possibility would be to minimize the corresponding Lagrangian dual. However, as pointed out in [? ?], this is computationally intractable for semantic segmentation networks involving millions of parameters; one has to optimize a CNN within each dual iteration. In fact, constrained optimization has been largely avoided in deep networks [?], even thought some Lagrangian techniques were applied to neural networks a long time before the deep learning era [? ?]. These constrained optimization techniques are not applicable to deep CNNs as they solve large linear system of equations[1].

[1] The numerical solvers underlying these constrained techniques would have to deal with matrices of very large dimensions in the case of deep networks [?].

In fact, to the best of our knowledge, the method of Pathak et al. [**?**] is the only prior work that addresses constrained deep CNNs in weakly supervised segmentation. It uses the constraints to synthesize fully-labeled training masks (proposals) from the available partial labels, mimicking full supervision, which avoids intractable dual optimization of the constraints when minimizing the loss function. The main idea of [**?**] is to model the proposals as a latent distribution. Then, they minimize a KL divergence, encouraging the softmax output of the CNN to match the latent distribution as closely as possible. Therefore, the impose constraints on the latent distribution instead of the network output, which facilitates significantly the optimization problem. We alternate between optimizing w.r.t the latent distribution, which corresponds to proposal generation subject to the constraints[2], and standard stochastic gradient descent for optimizing w.r.t the network parameters.

1.1 Contributions

We propose to introduce a differentiable term, which enforces inequality constraints (1) directly in the loss function, avoiding expensive Lagrangian dual iterates and proposal generation. From constrained optimization perspective, our simple approach is not optimal as there is no guarantee that the constraints are satisfied. However, surprisingly, it yields *substantially* better results than the proposal-based constrained CNNs in [**?**], while reducing the computational demand for training. In the context of cardiac image segmentation, we reached *a performance close to full supervision while using a fraction of the ground-truth labels (*0.1% *of the pixels of the ground-truth masks) with image-level tags*. Our framework can be easily extended to other inequality constraints, e.g., shape moments [**?**] or region statistics [**?**]. Therefore, it has the potential to close the gap between weakly and fully supervised learning in semantic medical image segmentation. Our code is publicly available[3].

2 Proposed loss function

We propose the following loss for weakly supervised segmentation:

$$\sum_{p\in\Omega} Y_p \log(S_p) + \mathcal{C}(S), \tag{2}$$

with function $\mathcal{C}$ given by (See the illustration in Fig. 1):

$$\mathcal{C}(S) = \begin{cases} \left(\sum_{p\in\Omega} S_p - a\right)^2, & \text{if } \sum_{p\in\Omega} S_p \leq a \\ \left(\sum_{p\in\Omega} S_p - b\right)^2, & \text{if } \sum_{p\in\Omega} S_p \geq b \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Now, our differentiable term $\mathcal{C}$ accommodates standard stochastic gradient descent. During back-propagation, the term of gradient-descent update corresponding to $\mathcal{C}$ can be written as follows:

$$-\frac{\partial \mathcal{C}(S)}{\partial \boldsymbol{\theta}} \propto \begin{cases} \left(a - \sum_{p\in\Omega} S_p\right) \frac{\partial S_p}{\partial \boldsymbol{\theta}}, & \text{if } \sum_{p\in\Omega} S_p < a \\ \left(b - \sum_{p\in\Omega} S_p\right) \frac{\partial S_p}{\partial \boldsymbol{\theta}}, & \text{if } \sum_{p\in\Omega} S_p > b \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\frac{\partial S_p}{\partial \boldsymbol{\theta}}$ denotes the standard derivative of the softmax outputs of the network. The gradient in (4) has a clear interpretation. During back-propagation, when the current constraints are satisfied, i.e., $a \leq \sum_{p\in\Omega} S_p \leq b$, observe that $\frac{\partial \mathcal{C}(S)}{\partial \boldsymbol{\theta}} = 0$. Therefore, in this case, the gradient stemming from our term has no effect on the current parameter update. Now, suppose without loss of generality that the current set of parameters $\boldsymbol{\theta}$ corresponds to $\sum_{p\in\Omega} S_p < a$, which means the current target region is smaller than its lower bound a (constraint violation). In this case, term $(a - \sum_{p\in\Omega} S_p)$ is positive and, therefore, the first line of (4) performs a gradient *ascent* step on softmax outputs, increasing S_p. This makes sense because it increases the size of the current region, $\sum_{p\in\Omega} S_p$, so as to satisfy the constraint. The case $\sum_{p\in\Omega} S_p > b$ has a similar interpretation.

[2] This sub-problem is convex when the constraints are convex.

[3]

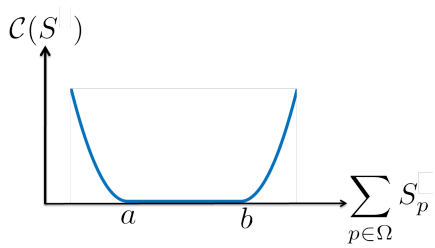

Figure 1: Illustration of our differentiable loss for imposing soft size constraints on the target region.

3 Experiments

In this section, we first give details about the dataset, the weakly annotated labels and our implementation. Then, we empirically evaluate our proposed high-order size loss. Particularly, we analyze its contribution to the segmentation performance, and compare the results to to the weakly supervised case with no regularization term, to image proposals generation and to the fully supervised setting.

3.1 Experimental set-up

3.1.1 Dataset

Experiments on the proposed high-order size loss focused on left ventricular endocardium segmentation. For this purpose we employed the training set from publicly available dataset from the 2017 ACDC Challenge [4]. This set consists on 100 cine magnetic resonance (MR) exams covering well defined pathologies: dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction with altered left ventricular ejection fraction, abnormal right ventricle and patients without cardiac disease. Exams were acquired in breath hold with a retrospective or prospective gating and with a SSFP sequence in 2-chambers, 4-chambers and in short axis orientations. A series of short axis slices cover the LV from the base to the apex, with a thickness of 5 to 8 mm and often an interslice gap of 5 mm. The spatial resolution goes from 0.83 to 1.75 mm^2/pixel.

For all the experiments, we employed the same 75 exams for training and the remaining 25 for validation purposes. To increase the variability of the data, we augment the dataset by randomly rotating, flipping, mirroring and scaling the images.

3.1.2 Weakly annotated labels

To generate the weak labels we employed binary dilation on the fully annotated labels with a kernel of size 10×10. If the resulted label disappeared, we repeated the operation with a smaller kernel (i.e., 7×7) until we get a small contour. Thus, the total number of annotated pixels represented the 0.1% of the labeled pixels in the fully supervised scenario. Figure 2 depicts some examples of fully annotated images and their corresponding weak labels.

To compute the lower and upper bounds of the proposed size loss, manual segmentations from only one subject were employed. Specifically, we computed the minimum and maximum size of left ventricular endocardium over the slices, and then multiplied by a factor of 1.1 and 0.9 the minimum and maximum value, respectively, to account for size variations across exams.

3.1.3 Training and implementation details

For all the proposed experiments we employ ENet [**?**] network, as it has shown a good trade-off between accuracy and inference time. Nevertheless, the proposed high-order size loss is general and it can be applied on any CNN. The network is trained from scratch by employing Adam optimizer and a batch size of 1. The initial learning rate was set to $5e{-}3$ and it was decreased by 2 after 100 epochs. The weight of the size loss in (4) was empirically set to $1e{-}2$ and its influence on the performance has not been investigated in this work. Input images had a size of 256×256 pixels.

[4] https://www.creatis.insa-lyon.fr/Challenge/acdc/

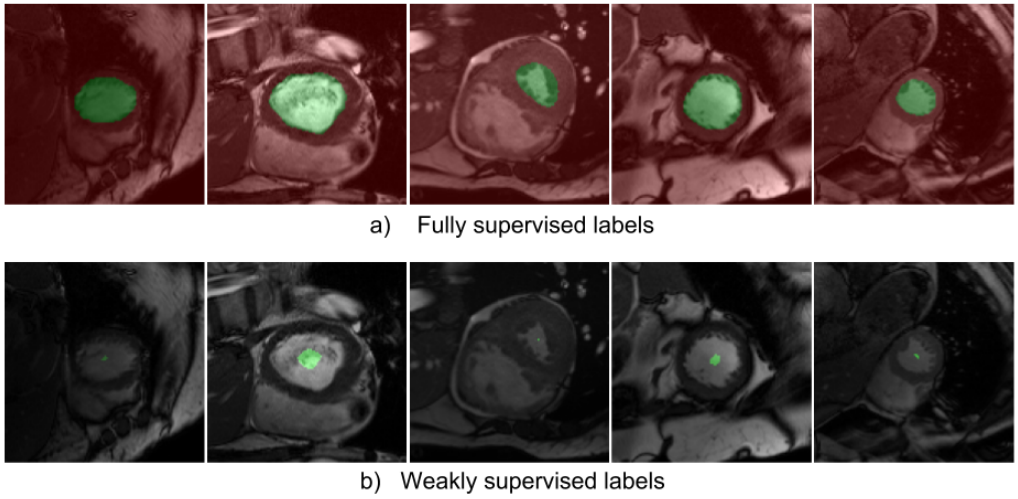

Figure 2: Examples of diverse supervisions. In fully labeled images (*top*) all pixels are annotated (Red pixels denote background whereas green pixels represent the object of interest). On weakly supervised cases (*bottom*), only labels of green pixels are known. Images have been cropped to better visualize the weak labels. Original images have a size of 256 $\times$ 256 pixels.

The proposals method [**?**] reuses the same network and loss function than the fully supervised setting. At each iteration, a synthetic ground truth $\tilde{Y}$ is generated using a projected gradient ascent, before computing the cross entropy between $\tilde{Y}$ and S. We found that limiting the number of iterations for the PGA to 500 (instead of the original 3000) saved time without impacting significantly the results. We employ one constraint ($\sum_p S^k < 5 * b$), with b following the same strategy as in Section 3.1.2. We had to add this multiplicative constant in order to make it work. This is discussed more in depth in section 3.2.2.

For evaluation purposes, since each method has a different loss function, we resort to the common Dice index $= \frac{2|S \bigcap Y|}{|S||Y|}$ to compare them.

We used a combination of PyTorch [**?**] and NumPy for our implementations, and ran the experiments on a machine equipped with a NVIDIA GTX 1080 Ti GPU (11GBs of memory). The code is available at https://github.com/LIVIAETS/SizeLoss_WSS.

3.2 Results

This section introduces the experimental results of this paper. First, in Sec. 3.2.1 we evaluate the impact of including an additional size loss during training in a weakly supervised setting. We show that by incorporating this size prior directly into the loss yields a highly significant boost on performance. Then, in Sec. 3.2.2 we compare the effect of using direct loss and generating proposals in training in an iterative scheme. Compared to this strategy, our proposed method achieved state-of-the-art performance on weakly supervised segmentation with small blobby regions as labels. We also provide the results for the fully supervised setting in Sec. 3.2.3. After that, qualitative evaluation is assessed in Sec. 3.2.4. In addition, we compare the different learning strategies in terms of efficiency (Sec. 3.2.5), showing that the direct introduction of the proposed loss during learning does not affect training times.

3.2.1 Weakly supervised segmentation with size loss

We trained a segmentation network from weakly annotated images with no additional information which served as baseline. Training on this model relies on computing the cross-entropy on labeled pixels. Then, we trained another network by employing the same weakly annotated images and we considered slightly more supervision than simply few annotated pixels. First, we included size prior information on the superior bound when the target is present in the image. In this case, a in

eq. 3 and 4 is replaced by 0. Thus, the CNN is constrained to generate segmentations whose sizes must be below the superior bound b. Then, if the target is not present on the image, we basically replace the upper bound b by 0, similar to MIL scenarios for suppression constraints. This will result on minimizing the predicted size to 0. To investigate the effect of bounding the CNN output on both sides, we trained a third network from weakly annotated images and we included size prior information as both inferior and superior bounds. Forward and backward pass are defined by eq. 3 and 4, where a and b are computed from one training subject following the steps described in Sec 3.1.2. The lower and upper size *tight* bounds have a value of 98 and 1723 pixels, respectively. Cross-entropy on annotated pixels and the proposed size loss are jointly minimized.

Table 1 reports results on the validation set for all the methods. The baseline is the network with cross entropy loss only on labeled pixels. This baseline achieves poor results, since no other information rather than the annotated pixels is given. As such, there is no regularization term that can give some hints on the desired segmentation. Evaluation on this baseline reported mean DSC values below 0.10, which are considered unsatisfactory. Indeed, after 100 epochs approximately (Fig. 3), we observed that the training got stuck and the network did not learn afterwards. On the other hand, constraining the CNN prediction with the proposed size loss significantly increased the performance of the network, as can be seen on both Table 1 and Fig 3. If we employ only a superior bound b, we achieve a mean DSC value of 0.8189. If a lower bound a is also considered, mean DSC increases up to 0.8415.

	Method	DSC (Val)
Weakly supervised	Cross-Entropy	0.0721
	Proposals (one bound)[1]	0.0659
	Proposals (one bound)[2]	0.6124
	Cross-Entropy + Size loss (one bound)[1]	0.8107
	Cross-Entropy + Size loss (one bound)[2]	0.8189
	Cross-Entropy + Size loss (two bounds)	**0.8415**
Fully supervised	—	0.9284

[1]Loose bound / [2]Tight bound

Table 1: Weakly supervised segmentation results for the diverse supervised learning strategies.

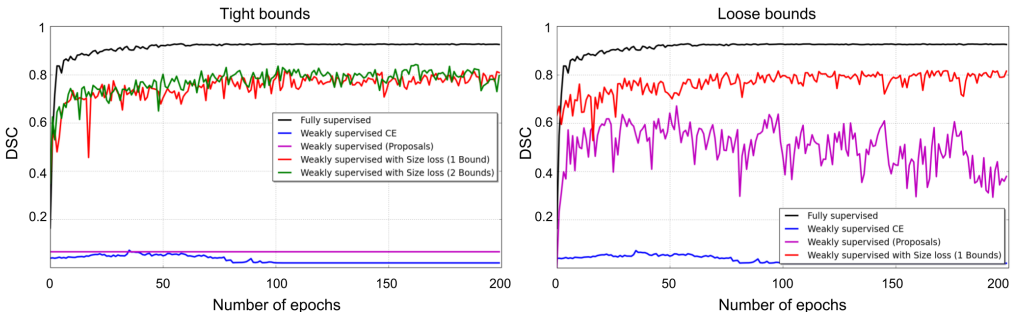

Figure 3: Evolution of DSC on validation samples for the weakly supervised learning strategies analyzed and fully supervision. We analyze difference on performances when employing tight versus loose bounds to constraint the size predictions.

3.2.2 Implicit loss against segmentation proposals

In this section we compare the implicit inclusion of the proposed high-order size loss during training to segmentation proposals generation [**?**]. First, we imposed the same *tight* bound employed in our proposed size loss. However, the performance of this method was deficient, as the network quickly produces only empty segmentations, and could not recover from them. This explains the flat line in the left plot in Fig. 3. We observed that by employing a much more *loose* upper bound – one sixth of the image size – this approach achieved better results, as reported on Table 1 and Fig. 3 (*right*). Learning of this method remained very unstable, while slowing overfitting. To compare fairly with

the proposed size loss, we also evaluated our loss with a *loose* bound equal to 10000. As shown on Table 1, constraining the learning with our size loss on a high bound reported an increase on performance with respect to the proposals generation approach in [**?**] under the same conditions.

Since we were not able to use tight bound on [**?**], we did not investigate the case with two bounds. Indeed, a single upper bound is equivalent to $0 < \sum_p S_p < b$, so adding a slightly over 0 – as in Sec. 3.2.1– would be effectively the same. As it can be seen on the results, compared to our direct size loss, generating proposals gives inferior segmentation performance on different situations.

3.2.3 Pixel-level annotations

Finally, we compared the weakly supervised strategies to a network trained on fully labeled images. Training the network with strong pixel-level annotations achieved a mean DSC of 0.9284 on the validation set. This result represents an increase of a 10% with respect to the best proposed method. Nevertheless, it is noteworthy to mention that the performance achieved by the proposed weakly supervised learning approach with size loss constraints approaches the fully supervised setting with only 0.1 $\%$ of the annotated pixels. This indicates that the current work shortens the gap between fully and weakly supervised learning for semantic segmentation, pushing the bounds of the latter.

3.2.4 Qualitative results

To get some intuition about the different learning strategies and their effects on the segmentation we visualize some results in Fig. 4. For the direct size loss and proposals generation approaches we selected the best performing model to display these images. We can observe that in all cases, the baseline gives unacceptable segmentations, which is in line with the findings observed in Table 1 and Fig. 3. We find that generating proposals during training might actually improve the segmentation performance compared to the baseline. Nevertheless, looking at the examples on Fig. 4 (*fourth column*) we can observe that these segmentations are far from being satisfactory. Nevertheless, integrating the proposed size loss directly in the back-propagation increases substantially the accuracy of the network, as can be seen in the last column of Fig. 4. An interesting observation is that, in some cases (*last row*), weakly supervised learning produces more reliable segmentations than training with full supervision.

3.2.5 Efficiency

In this section we compare the several learning approaches in terms of efficiency (Table 2). The weakly supervised baseline and the fully supervised model need to compute only one loss per pass. This is reflected in the lowest training times reported in the table. Including the size loss does not affect on the computation time, as can be seen in these results. As expected, the iterative process introduced by [**?**] at each forward pass cause a non negligible overhead during training. To generate their synthetic ground truth, they need to optimize the Lagragian function of their constrained loss with respect to its λ-parameters, which requires alternating between computing softmax probabilities and gradients of those probabilities. Even in simple optimization case (with only one constraint) where this parameter converge fast, it adds a few milliseconds at each iteration.

Method	Training time (ms/image)
Weakly supervised (CE Only)	10
Weakly supervised (CE + Size loss)	10
Weakly supervised (Proposals)	15
Fully supervised	10

Table 2: Training times for the diverse supervised learning strategies.

4 Conclusion

We presented a novel loss function for weakly supervised image segmentation, which despite its simplicity reach almost full supervision performances. We perform significantly better than the other proposed methods for this task, while having negligible computation overhead.

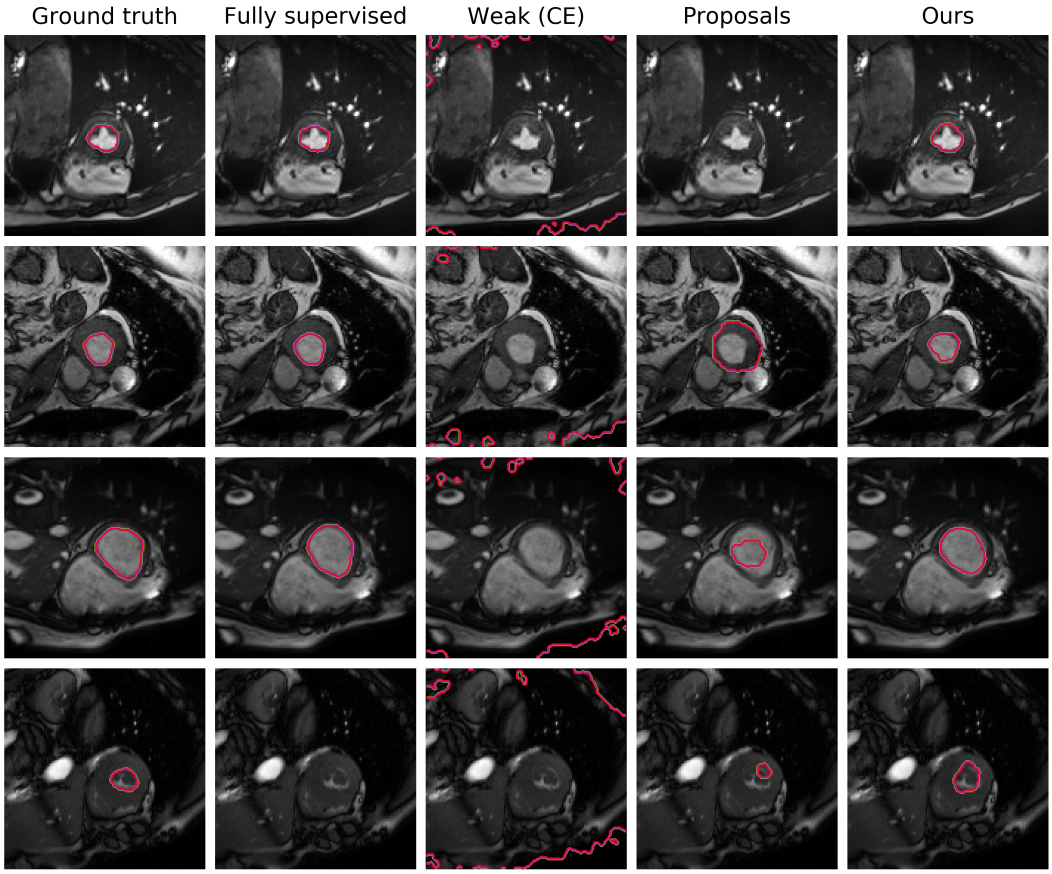

Figure 4: Qualitative comparison of the different methods. We can easily see that our method achieves results comparable to full supervision. Proposals fare good most of the time, but has a tendency to produce empty segmentations. Weak supervision with cross-entropy only quickly expand to the whole body. (Best viewed in colors)

Acknowledgments

This work is supported by the National Science and Engineering Research Council of Canada (NSERC), discovery grant program, and by the ETS Research Chair on Artificial Intelligence in Medical Imaging.