Constrained-CNN Losses for Weakly Supervised Segmentation

Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, Ismail Ben Ayed

PII: S1361-8415(18)30614-5
DOI: https://doi.org/10.1016/j.media.2019.02.009
Reference: MEDIMA 1464

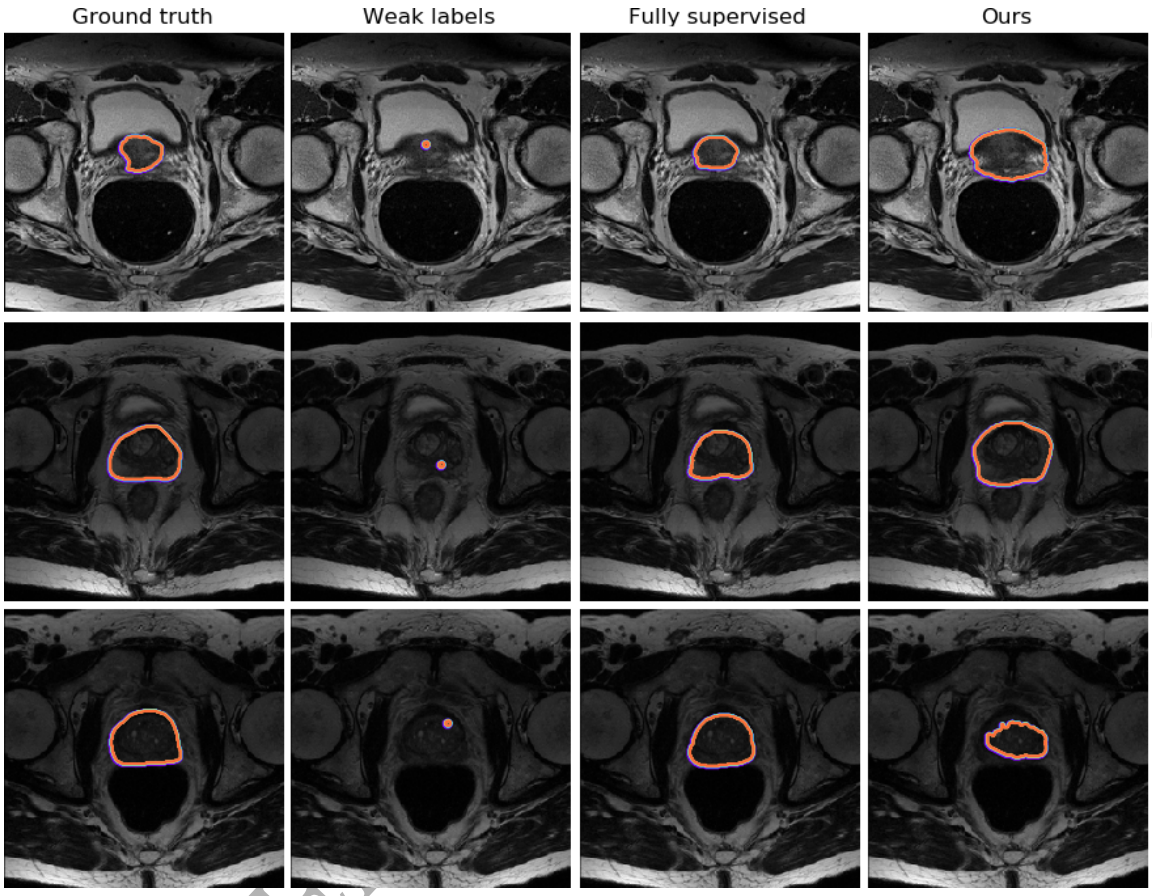
To appear in: *Medical Image Analysis*

Received date: 17 August 2018
Revised date: 4 February 2019
Accepted date: 12 February 2019

Please cite this article as: Hoel Kervadec, Jose Dolz, Meng Tang, Eric Granger, Yuri Boykov, Ismail Ben Ayed, Constrained-CNN Losses for Weakly Supervised Segmentation, *Medical Image Analysis* (2019), doi: https://doi.org/10.1016/j.media.2019.02.009

Highlights

- Impose **inequality constraints** during training of neural networks for semantic segmentation
- Constraints are **based on anatomical priors** (shape, size, ...)
- **Enable the use of weak annotations** such as dots, scribbles
- Add a penalty directly into the loss function, which is **very simple and efficient**
- Demonstrate the effectiveness on three **applications** (left-ventricle, vertebral-body, prostate segmentation
- Get **close to full supervision** performances with **0.1% of labeled pixels Code is publicly available**



Constrained-CNN Losses for Weakly Supervised Segmentation

Hoel Kervadec[a,*], Jose Dolz[a], Meng Tang[b], Eric Granger[a], Yuri Boykov[b], Ismail Ben Ayed[a]

[a]LIVIA, *ÉTS Montréal, QC, Canada*
[b]*Department of computer science, University of Waterloo, ON, Canada*

Abstract

Weakly-supervised learning based on, e.g., partially labelled images or image-tags, is currently attracting significant attention in CNN segmentation as it can mitigate the need for full and laborious pixel/voxel annotations. Enforcing high-order (global) inequality constraints on the network output (for instance, to constrain the size of the target region) can leverage unlabeled data, guiding the training process with domain-specific knowledge. Inequality constraints are very flexible because they do not assume exact prior knowledge. However, constrained Lagrangian dual 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. [1] is the only prior work that addresses deep CNNs with linear constraints 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 penalty, which enforces inequality constraints directly in the loss function, avoiding expensive Lagrangian dual iterates and proposal generation. From constrained-optimization perspective, our simple penalty-based approach is not optimal as there is no guarantee that the constraints are satisfied. However, surprisingly, it yields *substantially* better results than the Lagrangian-based constrained CNNs in [1], while reducing the computational demand for training. By annotating only a small fraction of the pixels, the proposed approach can reach a level of segmentation performance that is comparable to full supervision on three separate tasks. While our experiments focused on basic linear constraints such as the target-region size and image tags, our framework can be easily extended to other non-linear constraints, e.g., invariant shape moments [2] and other region statistics [3]. 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.

Keywords: Deep learning, semantic segmentation, weakly-supervised learning,

*Corresponding author: hoel.kervadec.1@etsmtl.net

CNN constraints.

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 [4, 5, 6]. 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. Furthermore, pixel/voxel level annotations become a serious impediment for scaling deep segmentation networks to new object categories or target domains.

To reduce the burden of pixel-level annotations, weak supervision in the form partial or uncertain labels, for instance, bounding boxes [7], points [8], scribbles [9, 10], or image tags [11, 12], is attracting significant research attention. Imposing prior knowledge on the network's output in the form of unsupervised loss terms is a well-established approach in machine learning [13, 14]. Such priors can be viewed as regularization terms that leverage unlabeled data, embedding domain-specific knowledge. For instance, the recent studies in [15, 10] 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. 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 [16, 17, 9, 18]. Typically, such proposal-based techniques iterate two steps: CNN learning and proposal generation facilitated by dense CRFs and fast mean-field inference [19], which are now the de-facto choice for pairwise regularization in semantic segmentation algorithms.

Our purpose here is to embed high-order (global) inequality constraints on the network outputs directly in the loss function, so as to guide learning. For instance, assume that we have some prior knowledge on the size (or volume) of the target region, e.g., in the form of lower and upper bounds on size, a common scenario in medical image segmentation [20, 21]. Let $I : \Omega \subset \mathbb{R}^{2,3} \rightarrow \mathbb{R}$ denotes a given training image, with Ω a discrete image domain and $|\Omega|$ the number of pixels/voxels in the image. $\Omega_L \subseteq \Omega$ is a weak (partial) ground-truth segmentation of the image, taking the form of a partial annotation of the target region, e.g., a few points (see Figure 2). In this case, one can optimize a *partial* cross-entropy loss subject to inequality constraints on the network outputs [1]:

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

where $\mathrm{S} = (S_1, \ldots, S_{|\Omega|}) \in [0, 1]^{|\Omega|}$ is a vector of softmax probabilities[1] generated

[1]The softmax probabilities take the form: $S_p(\boldsymbol{\theta}, I) \propto \exp f_p(\boldsymbol{\theta}, I)$, where $f_p(\boldsymbol{\theta}, I)$ is a real

by the network at each pixel p and $\mathcal{H}(S) = -\sum_{p\in\Omega_L} \log(S_p)$. 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 [22, 23, 24]. Also, multiple instance learning (MIL) constraints [1], which enforce image-tag priors, can be handled by constrained model (1). Image tags are a form of weak supervision, which enforce the constraints that a target region is present or absent in a given training image [1]. They can be viewed as particular cases of the inequality constraints in (1). For instance, a suppression constraint, which takes the form $\sum_{p\in\Omega} S_p \leqslant 0$, enforces that the target region is not in the image. $\sum_{p\in\Omega} S_p \geqslant 1$ enforces the presence of the region.

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 [1, 25], 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 [26], even thought some Lagrangian techniques were applied to neural networks a long time before the deep learning era [27, 28]. These constrained optimization techniques are not applicable to deep CNNs as they solve large linear systems of equations. The numerical solvers underlying these constrained techniques would have to deal with matrices of very large dimensions in the case of deep networks [25].

To the best of our knowledge, the method of Pathak et al. [1] is the only prior work that addresses inequality constraints in deep weakly supervised CNN 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 [1] is to model the proposals via a latent distribution. Then, it minimize a KL divergence, encouraging the softmax output of the CNN to match the latent distribution as closely as possible. Therefore, they impose constraints on the latent distribution rather than on the network output, which facilitates Lagrangian dual optimization. This decouples stochastic gradient descent learning of the network parameters and constrained optimization: The authors of [1] 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.

We propose to introduce a differentiable term, which enforces inequality constraints (1) directly in the loss function, avoiding expensive Lagrangian dual

scalar function representing the output of the network for pixel p. For notation simplicity, we omit the dependence of S_p on $\boldsymbol{\theta}$ and I as this does not result in any ambiguity in the presentation.

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

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 Lagrangian-based constrained CNNs in [1], 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 full ground-truth labels (0.1%). Our framework can be easily extended to non-linear inequality constraints, e.g., invariant shape moments [2] or other region statistics [3]. 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. Related work

2.1. Weak supervision for semantic image segmentation:

Training segmentation models with partial and/or uncertain annotations is a challenging problem [29, 30]. Due to the relatively easy task of providing global, image-level information about the presence or absence of objects in an image, many weakly supervised approaches used image tags to learn a segmentation model [31, 32]. For example, in [31], a probabilistic latent semantic analysis (PLSA) model was learned from image-level keywords. This model was later employed as a unary potential in a Markov random field (MRF) to capture the spatial 2D relationships between neighbours. Also, bounding boxes have become very popular as weak annotations due, in part, to the wide use of classical interactive segmentation approaches such as the very popular GrabCut [33]. This method learns two Gaussian mixture models (GMM) to model the foreground and background regions defined by the bounding box. To segment the image, appearance and smoothness are encoded in a binary MRF, for which exact inference via graph-cuts is possible, as the energies are sub-modular. Another popular form of weak supervision is the use of scribbles, which might be performed interactively by an annotator so as to correct the segmentation outcome.

GrabCut is a notable example in a wide body of "shallow" interactive segmentation works that used weak supervision before the deep learning era. More recently, within the computer vision community, there has been a substantial interest in leveraging weak annotations to train deep CNNs for color image segmentation using, for instance, image tags [1, 34, 35, 17, 11, 12], bounding boxes [7, 16, 36], scribbles [37, 9, 38, 15, 10] or points [8]. Most of these weakly supervised semantic segmentation techniques mimic full supervision by generating full training masks (segmentation proposals) from the weak labels. The proposals can be viewed as synthesized ground-truth used to train a CNN. In general,

[3]The code can be found at `https://github.com/LIVIAETS/SizeLoss_WSS`

these techniques follow an iterative process that alternates two steps: (1) standard stochastic gradient descent for training a CNN from the proposals; and (2) standard regularization-based segmentation, which yields the proposals. This second step typically uses a standard optimizer such mean-field inference [17, 16] or graph cuts [9]. In particular, the dense CRF regularizer of Krähenbühl and Koltun [19], facilitated by fast parallel mean-field inference, has become very popular in semantic segmentation, both in the fully [39, 40] and weakly [17, 16] supervised settings. This followed from the great success of DeepLab [40], which popularized the use of dense CRF and mean-field inference as a post-processing step in the context fully supervised CNN segmentation.

An important drawback of these proposal strategies is that they are vulnerable to errors in the proposals, which might reinforce themselves in such self-taught learning schemes [41], undermining convergence guarantee. The recent approaches in [15, 10] have integrated standard regularizers such as dense CRF or pairwise graph clustering directly into the loss functions, avoiding extra inference steps or proposal generation. Such direct regularization losses achieved state-of-the-art performances for weakly supervised color segmentation, reaching near full-supervision accuracy. While these approaches encourage pairwise consistencies between pixels during training, they do not explicitly impose global constraint as in (1).

2.2. Medical image segmentation with weak supervision:

Despite the increasing amount of works focusing on weakly supervised deep CNNs in semantic segmentation of color images, leveraging weak annotations in medical imaging settings is not simple. To our knowledge, the literature on this matter is still scarce, which makes weak-supervision approaches appealing in medical image segmentation. As in color images, common settings for weak annotations are bounding boxes. For instance, DeepCut [16] follows a similar setting as [17]. It generates image proposals, which are refined by a dense CRF before being re-used as "fake" labels to train the CNN. Using the bounding boxes as initializations for the Grab-cut algorithm, the authors showed that, by this iterative optimization scheme, one can obtain a performance better than the shallow counterpart, i.e., GrabCut. In another weakly supervised scenario [42], images were segmented in an unsupervised manner, generating a set of super-pixels [43], among which users had to select the regions belonging to the object of interest. Then, these masks generated from the super-pixels were employed to train a CNN. Nevertheless, as proposals are generated in an unsupervised manner, and due to the poor contrast and challenging targets typically present in medical images, these "fake" labels are likely prone to errors, which can be propagated during training, as stated before.

2.3. Constrained CNNs:

To the best of our knowledge, there are only a few recent works [1, 25, 24] that addressed imposing global constraints on deep CNNs. In fact, standard Lagrangian-dual optimization has been completely avoided in modern deep networks involving millions of parameters. As pointed out recently in [1, 25], there

is a consensus within the community that imposing constraints on the outputs of deep CNNs that are common in modern computer vision and medical image analysis problems is impractical: the direct use of Lagrangian-dual optimization for networks with millions of parameters requires training a whole CNN after each iterative dual step [1]. To avoid computationally intractable dual optimization, Pathak et al. [1] imposed inequality constraints on a latent distribution instead of the network output. This latent distribution describes a "fake" ground truth (or segmentation proposal). Then, they trained a single CNN so as to minimize the KL divergence between the network probability outputs and the latent distribution. This prior-art work is the most closely related to our study and, to our knowledge, is the only work that addressed inequality constraints in weakly supervised CNN segmentation. The work in [25] imposed hard equality constraints on 3D human pose estimation. To tackle the computational difficulty, they used a Kyrlov sub-space approach and limited the solver to only a randomly selected sub-set of the constraints within each iteration. Therefore, constraints that are satisfied at one iteration may not be satisfied at the next, which might explain the negative results in [25]. A surprising result in [25] is that replacing the equality constraints with simple L_2 penalties yields better results than Lagrangian optimization, although such a simple penalty-based formulation does not guarantee constraint satisfaction. A similar L_2 penalty was used in [24] to impose equality constraints on the size of the target regions in the context of histopathology segmentation. While the equality-constrained formulations in [25, 24] are very interesting, they assume exact knowledge of the target function (e.g., region size), unlike the inequality-constraint formulation in (1), which allows much more flexibility as to the required prior domain-specific knowledge.

3. Proposed loss function

We propose the following loss for weakly supervised segmentation:

$$\mathcal{H}(S) + \lambda \mathcal{C}\left(V_S\right), \tag{2}$$

where $V_S = \sum_{p \in \Omega} S_p$, λ is a positive constant that weighs the importance of constraints, and function $\mathcal{C}$ is given by (See the illustration in Fig. 1):

$$\mathcal{C}(V_S) = \begin{cases} \left(V_S - a\right)^2, & \text{if } V_S < a \\ \left(V_S - b\right)^2, & \text{if } V_S > b \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

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:

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

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., $a \leqslant V_S \leqslant b$, observe that $\frac{\partial \mathcal{C}(V_S)}{\partial \boldsymbol{\theta}} = 0$. Therefore, in this case, the gradient stemming from our term has no effect on the current update of the network parameters. Now, suppose without loss of generality that the current set of parameters $\boldsymbol{\theta}$ corresponds to $V_S < a$, which means the current target region is smaller than its lower bound a. In this case of constraint violation, term $(a - V_S)$ 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, V_S, so as to satisfy the constraint. The case $V_S > b$ has a similar interpretation.

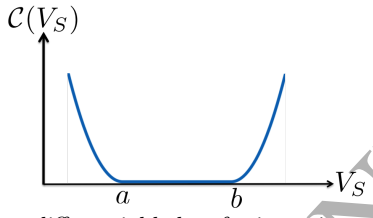


Figure 1: Illustration of our differentiable loss for imposing soft size constraints on the target region.

The next section details the dataset, the weak annotations and our implementation. Then, we report comprehensive evaluations of the effect of our constrained-CNN losses on segmentation performance. We also report comparisons to the Lagrangian-based constrained CNN method in [1] and to the fully supervised setting.

4. Experiments

4.1. Medical Image Data:

In this section, the proposed loss function is evaluated on three publicly available datasets, each corresponding to a different application – cardiac, vertebral body and prostate segmentation. Below are additional details of these data sets.

4.1.1. Left-ventricle (LV) on cine MRI

A part of our experiments focused on left ventricular endocardium segmentation. We used the training set from the publicly available data of the 2017 ACDC Challenge[4]. This set consists of 100 cine magnetic resonance (MR) exams covering well defined pathologies: dilated cardiomyopathy, hypertrophic cardiomyopathy, myocardial infarction with altered left ventricular ejection fraction and abnormal right ventricle. It also included normal subjects. Each exam

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

contains acquisitions only at the diastolic and systolic phases. The exams were acquired in breath-hold with a retrospective or prospective gating and 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 an inter-slice 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.

4.1.2. Vertebral body (VB) on MR-T2

This dataset contains 23 3D T2-weighted turbo spin echo MR images from 23 patients and the associated ground-truth segmentation, and is freely available from [5]. Each patient was scanned with 1.5 Tesla MRI Siemens scanner (Siemens Healthcare, Erlangen, Germany) to generate T2-weighted sagittal images. All the images are sampled to have the same sizes of $39\times305\times305$ voxels, with a voxel spacing of $2\times1.25\times1.25$ mm^3. In each image, 7 vertebral bodies, from T11 to L5, were manually identified and segmented, resulting in 161 labeled regions in total. For this dataset, we employed 15 scans for training and the remaining 5 for validation.

4.1.3. Prostate segmentation on MR-T2

The third dataset was made available at the MICCAI 2012 prostate MR segmentation challenge[6]. It contains the transversal T2-weighted MR images of 50 patients acquired at different centers with multiple MRI vendors and different scanning protocols. It is comprised of various diseases, i.e., benign and prostate cancers. The images resolution ranges from $15 \times 256 \times 256$ to $54 \times 512 \times 512$ voxels with a spacing ranging from $2 \times 0.27 \times 0.27$ to $4 \times 0.75 \times 0.75 mm^3$. We employed 40 patients for training and 10 for validation.

4.2. Weak annotations:

To show that the proposed approach is robust to the strategy for generating the weak labels, as well as to their location, we consider two different strategies generating weak annotations from fully labeled images. Figure 2 depicts some examples of fully annotated images and the corresponding weak labels.

Erosion. For the left-ventricle dataset, we employed binary erosion on the fully annotations 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. This correspond to the second row in Figure 2.

[5] http://dx.doi.org/10.5281/zenodo.22304
[6] `https://promise12.grand-challenge.org`

Random point. The weak labels for the vertebral body and prostate datasets were generated by randomly selecting a point within the ground-truth mask and creating a circle around it with a maximum radius of 4 pixels (fourth and sixth row in Fig. 2), while ensuring there is no overlap with the background. With these weak annotations, only 0.02% of the pixels in the dataset have ground-truth labels.

4.3. Different levels of supervision:

Training models with diverse levels of supervision requires that appropriate objectives be defined for each case. In this section, we introduce the different models, each with different levels of supervision.

4.3.1. Baselines

We trained a segmentation network from weakly annotated images with no additional information, which served as a lower baseline. Training this model relies on minimizing the cross-entropy corresponding to the fraction of labeled pixels: $\mathcal{H}(S) = -\sum_{p \in \Omega_L} \log(S_p)$. In the following discussion of the experiments, we refer to this model as *partial cross-entropy (CE)*.

As an upper baseline, we resort to the fully-supervised setting, where class labels (foreground and background) are known for every pixel during training ($\Omega_L = \Omega$). This model is referred to as *fully-supervised*.

4.3.2. Size constraints

We incorporated information about the size of the target region during training, and optimized the partial cross-entropy loss subject to inequality constraints of the general form in Eq. (1). We trained several models using the same weakly annotated images but different constraint values.

Image tags bounds. Similar to MIL scenarios, we first used image-tag priors by enforcing the presence or absence of a the target in a given training image, as introduced earlier. This reduces to enforcing that the size of the predicted region is less or equal to 0 if the target is absent from the image, or larger than 0 otherwise. To simplify the implementation, we can represent the constraints as:

$$a, b = \begin{cases} 1, |\Omega| & \text{if target is present } (\Omega_L \neq \varnothing) \\ 0, 0 & \text{otherwise} \end{cases}. \tag{5}$$

While being very coarse, these constraints convey relevant information about the target regions, which may be used to find common patterns in the case of region absence or presence.

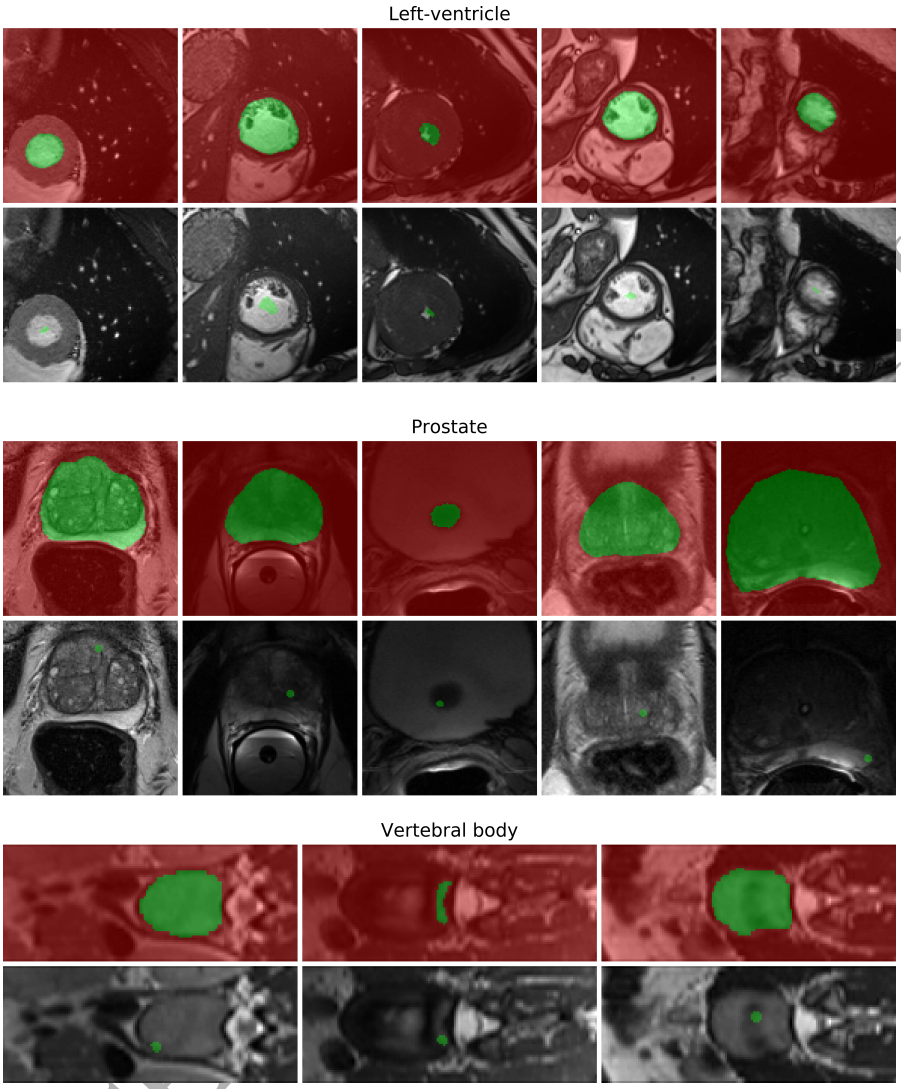


Figure 2: Examples of different levels of supervision. In the fully labeled images (*top*), all pixels are annotated, with red depicting the background and green the region of interest. In the weakly supervised cases (*bottom*), only the labels of the green pixels are known. The images were cropped for a better visualization of the weak labels. The original images are of size 256 × 256 pixels.

Common bounds. The next level of supervision consists of using tighter bounds for the positive cases, instead of $(1, |\Omega|)$. To this end, the complete segmentation of a *single* patient is employed to compute the minimum and maximum size of the target region across all the slices. Then, we multiplied these minimum

and maximum values by 0.9 and 1.1, respectively, to account for inter-patient variability. In this case, all the images containing the object of interest have the same lower and upper bounds. As an example, this results in the following values for the ACDC dataset:

$$a, b = \begin{cases} 60, 2000 & \text{if target is present } (\Omega_L \neq \varnothing) \\ 0, 0 & \text{otherwise} \end{cases}. \tag{6}$$

Individual bounds. With common bounds, the range of values for a given target may be very large. To investigate whether a more precise knowledge of the target is helpful, we also consider the use of individual bounds for each slice, based on the true size of the region:

$$\tau_Y = \sum_{p \in \Omega} Y_p,$$

with $Y = (Y_1, ..., Y_{|\Omega|}) \in \{0, 1\}^{|\Omega|}$ denoting the full annotation of image I. As before, we introduce some uncertainty on the target size, and multiply τ_Y by the same lower and upper factors, resulting in the following bounds:

$$a, b = \begin{cases} 0.9\tau_Y, 1.1\tau_Y & \text{if target is present } (\Omega_L \neq \varnothing) \\ 0, 0 & \text{otherwise} \end{cases}. \tag{7}$$

4.3.3. Hybrid training

We also investigate whether combining our proposed weak supervision approach with fully annotated images during the training leads to performance improvements. For this purpose, considering we have a training set of m weakly annotated images, we replace n ($n < m$) among these by their fully annotated counterparts. Thus, the training amounts to minimizing the cross-entropy loss for the n fully annotated images, along with the partial cross-entropy constrained with common size bounds for the remaining $m - n$ weakly labeled images. To examine the positive effect of size constraints in this scenario (referred to as *Hybrid*), we compare the results to a network trained with the n fully annotated images (without constraints).

4.4. Constraining a 3D volume:

We can extend our formulation to constrain a 3D volume as follows:

$$\sum_{S \in \mathrm{B}} \mathcal{H}(S) + \lambda \mathcal{C}(V_\mathrm{B}), \qquad \text{with} \quad V_\mathrm{B} = \sum_{S \in \mathrm{B}} V_S$$

where V_B denotes the target-region volume, $\mathrm{B} = ((Y^1, S^1), ..., (Y^{|B|}, S^{|B|}))$ denotes a training batch containing all the 2D slices of the 3D volume[7], and the

[7]For readability, we simplify a batch as a list of labels Y and associated predictions S.

3D constraints are now given by:

$$a, b = 0.9\tau_{\mathrm{B}}, 1.1\tau_{\mathrm{B}}, \qquad \text{with} \quad \tau_{\mathrm{B}} = \sum_{Y \in \mathrm{B}} \tau_Y$$

Notice that, with constraints on the whole 3D volume, we have less supervision than the 2D scenarios from 4.3.2, where all the 2D slices have independent supervision (e.g., the image tags).

4.5. Training and implementation details:

For the experiments on the left-ventricle and vertebral-body datasets, we used ENet [44], as it has shown a good trade-off between accuracy and inference time. Due to the higher difficulty of the prostate segmentation task, we employed a fully residual version of U-Net [45], similar to [46].

For the three datasets, we trained the networks from scratch using the Adam optimizer and an initial learning rate of 5×10^{-4} that we decreased by a factor of 2 if the performances on the validation set did not improve over 20 epochs. All the 3D volumes were sliced into 256×256 pixels images, and zero-padded when needed. Batch sizes were equal to 1, 4 and 20 for the left-ventricle, prostate and vertebral body, respectively. Those values were not tuned for optimal performances, but to speed-up experiments when enough data were available. The weight of our loss in (2) was empirically set to 1×10^{-2}. Due to the difficulty of the task, data augmentation was used for the prostate dataset, where we generated 4 copies of each training image using random mirroring, flipping and rotation.

All our tests were implemented in Pytorch [47]. We ran the experiments on a machine equipped with a NVIDIA GTX 1080 Ti GPU (11GBs of video memory), AMD Ryzen 1700X CPU and 32GBs of memory. The code is available at https://github.com/LIVIAETS/SizeLoss_WSS. We used the common Dice similarity coefficient (DSC) to evaluate the segmentation performance of trained models.

4.5.1. Modification and tweaks for Lagrangian proposals

For a fair comparison, we re-implemented the Lagrangian-proposal method of Pathak et al. [1] in PyTorch, to take advantage of GPU capabilities and avoid costly transfers between GPU and CPU. Lagrangian proposals reuse the same network and loss function as the fully-supervised setting. At each iteration, the method alternates between two steps. First, it synthesizes a ground truth $\tilde{Y}$ with projected gradient ascent (PGA) over the dual variables, with the network parameters fixed. Then, for fixed $\tilde{Y}$, the cross-entropy between $\tilde{Y}$ and S is optimized as in standard fully-supervised CNN training. The learning rate used for this PGA was set experimentally to 5×10^{-5}, as sub-optimal values lead to numerical errors. We found that limiting the number of iterations for the PGA to 500 (instead of the original 3000) saved time without affecting the results.

We also introduced an early stopping mechanism into the PGA in the case of convergence, to improve speed without impacting the results (a comparison can be found in Table 5). The constraints of the form $0 \leqslant V_S \leqslant 0$ required specific care, as the formulation from [1] is not designed to work on equalities, unlike our penalty approach, which systematically handles equality constraints when $a = b$. In this case, the bounds for [1] were modified to $-1 \leqslant V_S \leqslant 0$.

5. Results

To validate the proposed approach, we first performed a series of experiments focusing on LV segmentation. In Sec. 5.1, the impact of including size constraints is evaluated using our direct penalty. We further compare to the Lagrangian-proposal method in [1], showing that our simple method yields substantial improvements over [1] in the same weakly supervised settings. We also provide the results for several degrees of supervision, including hybrid and fully supervised learning in Sec. 5.2. Then, to show the wide applicability of the proposed constrained loss, results are reported for two other applications in Sec. 5.3: MR-T2 vertebral body segmentation and prostate segmentation task. We further provide qualitative results for the three applications in Sec. 5.4. In Sec. 5.5, we investigate the sensitivity of the proposed loss to both the lower and upper bounds. Finally, the efficiency of different learning strategies are compared (Sec. 5.6), showing that our direct constrained-CNN loss does not add to the training time, unlike the Lagrangian-proposal method in [1].

5.1. Weakly supervised segmentation with size constraints:

2D segmentation. Table 1 reports the results on the left-ventricle validation set for all the models trained with both the Lagrangian proposals in [1] and our direct loss. As expected, using the partial cross entropy with a fraction of the labeled pixels yielded poor results, with a mean DSC less than 15%. Enforcing the image-tag constraints, as in the MIL scenarios, increased substantially the DSC to a value of 0.7924. Using common bounds increased the results marginally in this case, slightly increasing the mean Dice value by 1%. The Lagrangian proposal [1] reaches similar results, albeit slightly lower and much more unstable than our penalty approach (see Figure 3).

The difference in performance is more pronounced when we employ individual bounds instead. In this setting, our method achieves a DSC of 0.8708, only 2% lower than full supervision. However, the Lagrangian-proposal method achieves a performance similar to using common (loose) bounds, suggesting that it is not able to make use of this extra, more precise information. This can be explained by its proposal-generation method, which tends to reinforce early mistakes (especially when training from scratch): the network is trained with conflicting information – i.e., similar-looking patches are both foreground and background according the the synthetic ground truth – and is not able to recover from those initial mis-classifications.

3D segmentation. Constraining the size of the 3D volume of the target region also shows the benefit of our penalty approach, yielding a mean DSC of 0.8580. Recall that, here, we are using less supervision than the 2D case. Since we do not use tag information in this case, these results suggest that only a fraction of all the slices may be used when creating the labels, allowing annotators to scribble the 3D image directly instead of going through all the 2D slices one by one.

Table 1: Left-ventricle segmentation results with different levels of supervision. Bold font highlights the best weakly supervised setting.

	Model	Method	DSC (Val)
Weakly supervised	Partial CE		0.1497
	CE + Tags	Lagrangian Proposals [1]	0.7707
	Partial CE + Tags	Direct loss (Ours)	0.7924
	CE + Tags + Size*	Lagrangian Proposals [1]	0.7854
	Partial CE + Tags + Size*	Direct loss (Ours)	0.8004
	CE + Tags + Size**	Lagrangian Proposals [1]	0.7900
	Partial CE + Tags + Size**	Direct loss (Ours)	**0.8708**
	CE + 3D Size**	Lagrangian Proposals [1]	N/A
	Partial CE + 3D Size**	Direct loss (Ours)	0.8580
Fully supervised	Cross-entropy		0.8872

*Common bounds / ** Individual bounds

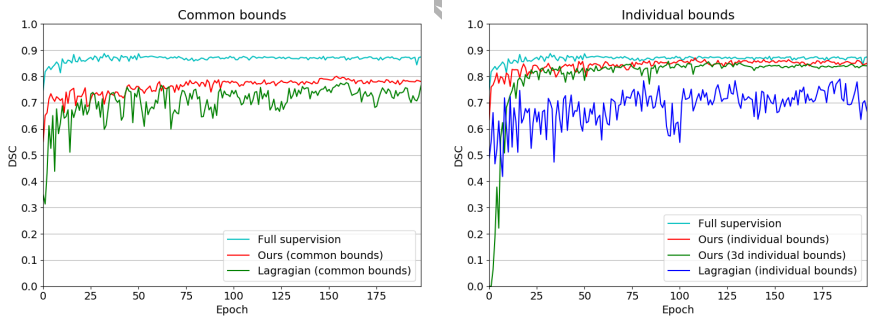


Figure 3: Evolution of the DSC during training for the left-ventricle validation set, including the weakly supervised learning models and different strategies analyzed, with also the full-supervision setting. As tags and common bounds achieve similar results, we plot only common bounds for better readability.

5.2. Hybrid training: mixing fully and weakly annotated images:

Table 2 and Figure 4 summarize the results obtained when combining weak and full supervision. First, and as expected, we can observe that adding n fully annotated images to the training set (Hybrid_n) improves the performances in comparison to the model trained solely with the weakly annotated images, i.e., Weak_All. Particularly, the DSC increases by 4%,5% and 6% when n is equal

to 5,10 and 25, respectively, approaching the full-supervision performance with only 25% of the fully labeled images.

Nevertheless, it is more interesting to see the impact of adding weakly annotated images (i.e., Hybrid_n) to a model trained solely with fully labeled images (i.e., Full_n). From the results, we can observe that adding weakly annotated images to the training set significantly increases the performance, particularly when the amount of fully annotated images (i.e., n) is limited. For instance, in the case of n equal to 5, adding weakly annotated images enhanced the performance by more than 30% in comparison to full supervision with n equal to 5. Despite the fact that this gap decreases with the number of fully annotated images, the difference between both settings (i.e., Full and Hybrid) remains significant. More interestingly, training the same model with a high amount of weakly annotated images and no or a very reduced set of fully labeled images (for example Weak_All or Hybrid_5) achieves better performances than employing datasets with much higher numbers of fully labeled images, e.g, Full_25.

These results suggest that a good strategy when annotating a new dataset might be to start with weak labels for all the images, and progressively complete full annotations, should ressources become available.

Table 2: Ablation study on the amounts of fully and weakly labeled data. We report the mean DSC of all the testing cases, for all the settings and using the same architecture.

Name	**Training approach**	**# Fully/Weakly annotated images**	**DSC**
Weak_All	Weak supervision*	0/150	0.8004
Full_5	Full supervision	5/0	0.5434
Hybrid_5	Full + weak supervision*	5/145	0.8386
Full_10	Full supervision	10/0	0.6004
Hybrid_10	Full + weak supervision*	10/140	0.8475
Full_25	Full supervision	25/0	0.7680
Hybrid_25	Full + weak supervision*	25/125	0.8641
Full_All	Full supervision	150/0	0.8872

*Common bounds

5.3. MR-T2 vertebral body and prostate segmentation:

The results obtained for the vertebral-body dataset (Table 3) highlight well the differences in the performances of different levels of supervision. Using tag bounds produces a network that roughly locates the object of interest (DSC of 0.5597), but fails to identify its boundaries (as seen in Figure 6, *third column*). Employing the common size strategy achieves satisfactory results for the slices containing objects with a regular shape but still fails when more difficult/irregular targets are present, resulting in an overall improvement of DSC (0.7900). However, when using individual bounds, the network is able to satis-

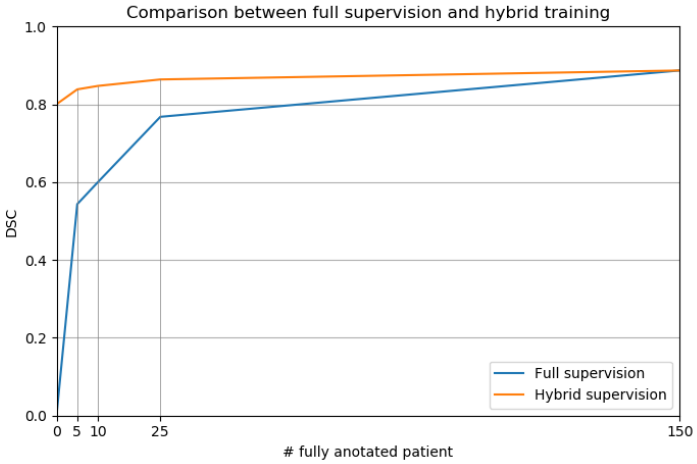


Figure 4: Mean DSC values over the number of fully annotated patients employed for training.

factory segment even the most difficult cases, obtaining a DSC of 0.8604, only 3% lower than full supervision.

Table 3: Mean Dice scores (DSC) for several degrees of supervision, using the vertebral-body and prostate validation sets. Bold font indicates the best weakly supervised setting for each data set.

Method	Vertebral body DSC	Prostate DSC
Partial CE	0.1155	0.0320
Partial CE + Tags	0.5597	0.6911
Partial CE + Tags + Common size	0.7900	0.7214
Partial CE + Tags + Individual size	**0.8604**	**0.8298**
Fully supervised	0.8999	0.8911

For the prostate dataset, one can observe that common bounds still improve the results obtained with tags (+3%), but the difference is much smaller than the case of vertebral-body segmentation. Using individual bounds increases the DSC value by 10%, reaching 0.8298, a behaviour similar to what we observed earlier for the other datasets. Nevertheless, in this case, the gap between full and weak supervision with individual bounds constraints is larger than what we obtained for the other datasets.

5.4. Qualitative results:

To gain some intuition on different learning strategies and their impact on the segmentation, we visualize some results sampled from the validation sets in Fig. 5, 6 and 7 for LV, VB and prostate, respectively.

LV segmentation task. We compare 4 methods to the ground truth: full supervision, Lagrangian proposals [1] with common bounds, direct loss with common bounds and direct loss with individual bounds. We can see that, for the easy cases containing regular shapes and visible borders, all methods obtain similar results. However, the methods employing common bounds can easily oversegment the object, especially when their size is considerably smaller; see for example the last row in Figure 5. Since individual bounds are specific to each image, a model trained with these bounds will not suffer in such cases, as shown in the figure.

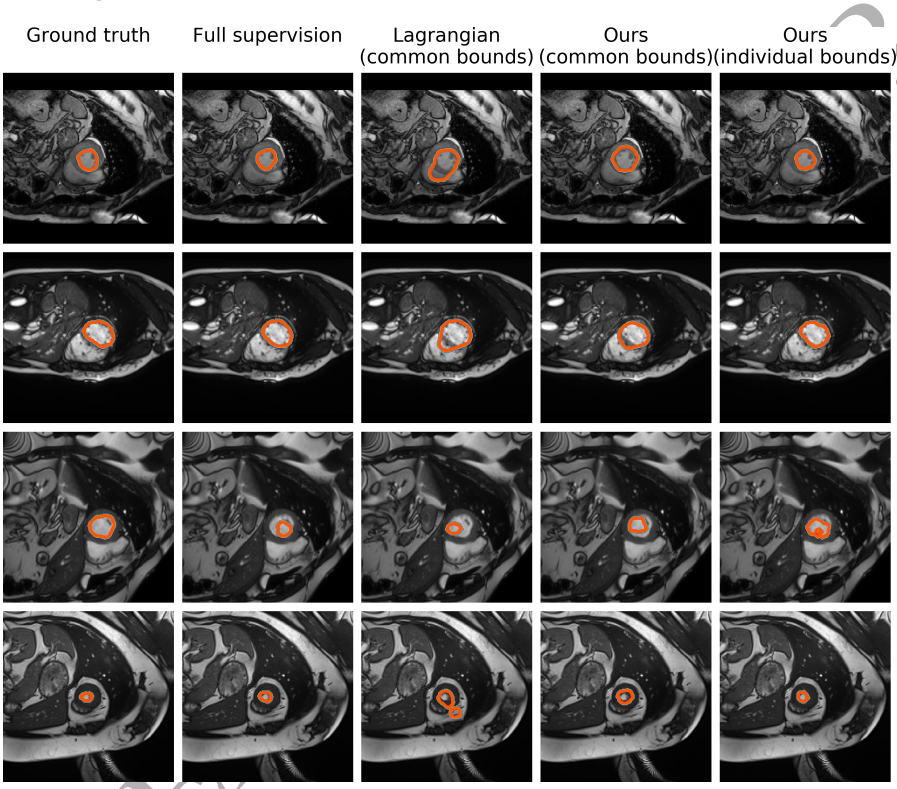


Figure 5: Qualitative comparison of the different methods using examples from the LV dataset. Each column depicts segmentations obtained by different methods, whereas each row represents a 2D slice from different scans (Best viewed in colors).

Vertebral-body segmentation task. In this case, we visualize the results of full supervision, tag bounds, common bounds and individual bounds. In line with results reported in Table 3, we can visually observe the gap in performances between each setting, which clearly highlights the impact of the different values of the bounds during the optimization process. Using only tags, the network learn to roughly locate the object. When size bounds are included as common

size information, the network is able to somehow learn the boundaries, but only for object shapes that are within the standard variability of a typical vertebral body shape. As it can be observed, the model fails to segment the unusual shapes (last three rows in Figure 6). Lastly, a network trained with individual sizes is able to better handle those cases, while still being imprecise on some regions.

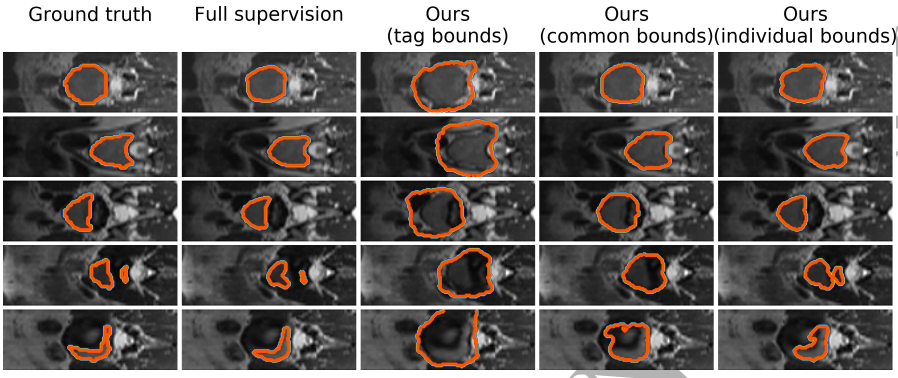


Figure 6: Qualitative comparison using examples from the VB dataset. Each column depicts segmentations obtained by different levels of supervision, whereas each row represents a 2D slice from different scans (Best viewed in colors).

Prostate segmentation task. As in the previous case, we depict the results of full supervision, tag bounds, common bounds and individual bounds. Both the tags and common bounds locate the object in a similar fashion, but both have difficulties finding a precise contour, typically over-segmenting the target region. This is easily explained by the variability of the organ and the very low contrast on some images. As shown in the last column, using individual bounds greatly improves the results.

5.5. Sensitivity to the constraint boundaries:

In this section, an ablation study is performed on the lower and upper bounds when using common bounds, and investigate their effect on the performance on the vertebral-body segmentation task. Results for different bounds are reported in Table 4. It can be observed that progressively increasing the value of the upper bound decreases the performance. For example, the DSC drops by nearly 12% and 16% when the upper bound is increased by a factor of 5 and 10, respectively. Decreasing the lower bound from 80 to 0 has a much smaller impact than the upper bound, with a constant drop of less than 1%. These findings are aligned with visual predictions illustrated in Figure 6. While a network trained only with tag bounds tends to over-segment, adding an upper bound easily fixes the over-segmentation, correcting most of the mistakes. Nevertheless, for the same reason, i.e., over-segmentation, very few slices benefit from a lower bound.

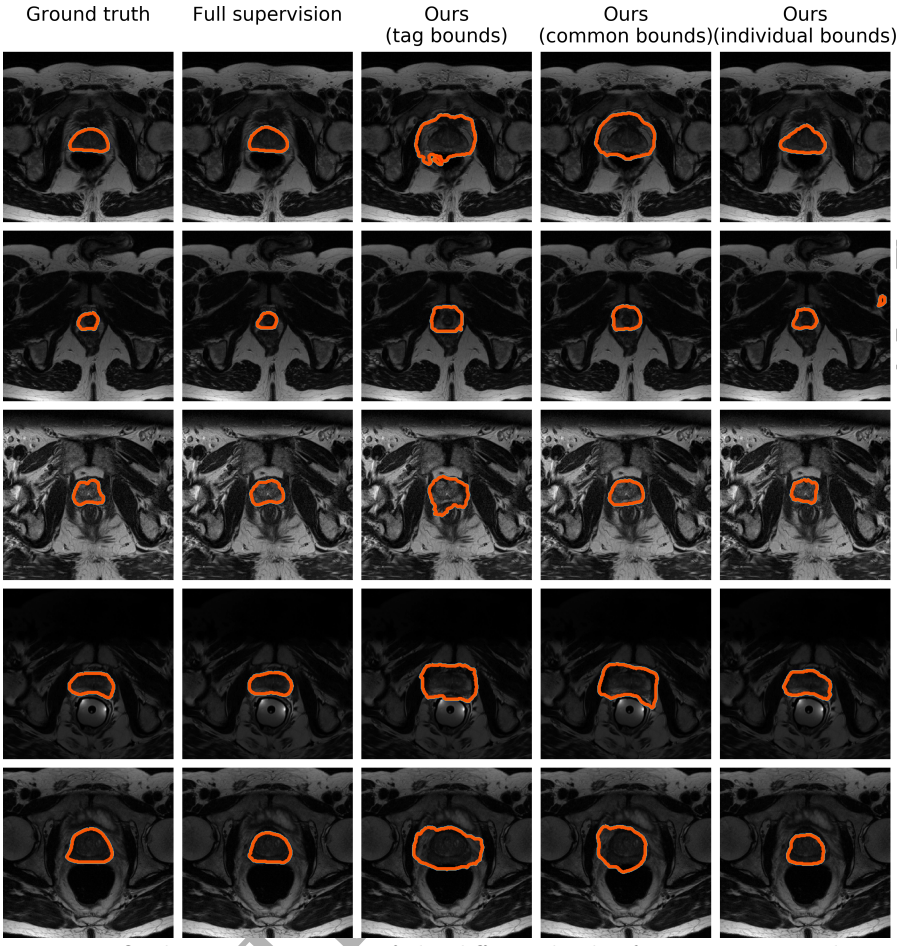


Figure 7: Qualitative comparison of the different levels of supervision. Each row represents a 2D slice from different scans. (Best viewed in colors)

5.6. Efficiency:

In this section, we compare the several learning approaches in terms of efficiency (Table 5). Both the weakly supervised partial cross-entropy 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 add to the computational time, as can be seen in these results. As expected, the iterative process introduced by [1] at each forward pass adds a significant overhead during training. To generate their synthetic ground truth, they need to optimize the Lagrangian function with respect to its dual variables (Lagrange multipliers of the constraints), which requires alternating between training a CNN and Lagrangian-dual optimization. Even in the simplest optimization case (with only one constraint), where optimization over the dual

Table 4: Ablation study on the lower and upper bounds of the size constraint using the vertebral body dataset.

	Bounds		**Mean DSC**
Model	Lower (a)	Upper (b)	
Weak Sup. w/ direct loss	$0.9\tau_Y$	$1.1\tau_Y$	0.8604
Weak Sup. w/ direct loss	80	1100	0.7900
Weak Sup. w/ direct loss	80	5000	0.6704
Weak Sup. w/ direct loss	80	10000	0.6349
Weak Sup. w/ direct loss	0	1100	0.7820
Weak Sup. w/ direct loss	0	5000	0.6694
Weak Sup. w/ direct loss	0	10000	0.6255
Weak Sup. w/ direct loss	0	65536	0.5597

variable converges rapidly, their method remains two times slower than ours. Without the early stopping criteria that we introduced, the overhead is much worse with a six-fold slowdown. In addition, their method also slows down when more constraints are added. This is particularly significant when there is many classes to constrain/supervise.

Generating the proposals at each iteration also makes it much more difficult to build an efficient implementation for larger batch sizes. One either needs to generate them one by one (so the overhead grows linearly with the batch size) or try to perform it in parallel. However, due to the nature of GPU design, the parallel Lagrangian optimizations will slow each other down, meaning that there may be limited improvements over a sequential generation. In some cases it may be faster to perform it on CPU (where the cores can truly perform independent tasks in parallel), at the cost of slow transfers between GPU and CPU. The optimal strategy would depend on the batch size and the host machine, especially its available GPU, number of CPU cores and bus frequency.

Table 5: Training times for the diverse supervised learning strategies with a batch size of 1, using tags and size constraints.

Method	Training time (ms/batch)
Partial CE	112
Direct loss (1 bound)	113
Direct loss (2 bounds)	113
Lagrangian proposals (1 bound)	610
Lagrangian proposals (2 bounds)	675
Lagrangian proposals (1 bound), w/ early stop	221
Lagrangian proposals (2 bounds), w/ early stop	220
Fully supervised	112

6. Discussion

We have presented a method to train deep CNNs with linear constraints in weakly supervised segmentation. To this end, we introduce a differentiable term, which enforces inequality constraints directly in the loss function, avoiding expensive Lagrangian dual iterates and proposal generation.

Results have demonstrated that leveraging the power of weakly annotated data with the proposed direct size loss is highly beneficial, particularly when limited full annotated data is available. This could be explained by the fact that the network is already trained properly when a large fully annotated training set is available, which is in line with the values reported in Table 2. Similar findings were reported in [48, 49], where authors exhibited an increased of performance when including non-annotated images in a semi-supervised setting. This suggests that including more unlabelled or weakly labelled data can potentially lead to significantly improvements in performance.

Findings from experiments across different segmentation tasks indicate that highly competitive performance can be obtained with a rough estimation of the target size. This is especially the case on well structured problems where the size and/or shape of the object remains consistent across subjects. If more precise size bounds are provided, the proposed approach is able to reach performances close to full supervision, even when the size and shape variability across subjects is large. For difficult tasks, where the gap between our approach and full supervision is larger, such as prostate segmentation, including an unsupervised regularization loss [10, 15] to encourage pairwise consistencies between pixels may boost the performance of the proposed strategy. A noteworthy point is the robustness of our method to the weak-label generation. While the weak labels were generated from a ground-truth erosion for the first dataset, with seeds always in the center of the target region, they were randomly generated and placed for the other two datasets. Thus, the results showed consistency in the behaviour of the different methods, regardless of the strategy used.

Even though the proposed method has been shown to provide good generalization capabilities across three different applications, the segmentation of images with severe abnormalities, whose sizes largely differ from those seen in the training set, has not been assessed. Nevertheless, the ablation study performed on the values of the size bounds, and the results obtained with common bound sizes suggest that the proposed approach may perform satisfactorily in the presence of these severe abnormalities, by simply increasing the upper bound value. In addition, if a greater 'precise' estimation of the abnormality size is given, our proposed loss may improve segmentation performance, as demonstrated by the results achieved by the individual bounds strategy. It is important to note that, even in the case of full supervision, if a new testing image contains a severe abnormality much larger than the objects seen during the training phase, the network will likely to poorly segment the region of interest.

Our framework can be easily extended to other non-linear (fractional) constraints, e.g., invariant shape moments [2] or other statistics such as the mean of intensities within the target regions [3]. For instance, a normalized (scale

invariant) shape moment of a target region can be directly expressed in term of network outputs using the following general fractional form:

$$F_S = \frac{\sum_{p\in\Omega} f_p S_p}{\sum_{p\in\Omega} S_p} \tag{8}$$

where f_p is a unary potential expressed in term of exponents of pixel/voxel coordinates. For example, the coordinates of the center of mass of the target region are particular cases of (8) and correspond to first-order scale-invariant shape moments. In this case, potentials f_p correspond to pixel coordinates. Now, assume a weak-supervision scenario in which we have a rough localization of the centroid of the target region. In this case, instead of a constraint on size representation V_S as in Eq. (3), one can use a cue on centroid as follows: $a \leqslant F_S \leqslant b$. This can be embedded as a direct loss using differentiable penalty $\mathcal{C}(F_S)$. Of course, here, F_S is a non-linear fractional term unlike region size. Therefore, in future work, it would be interesting to examine the behaviour of such fractional terms for constraining deep CNNs with a penalty approach. Finally, it is worth noting that the general form in Eq. (8) is not confined to shape moments. For instance, the image (intensity) statistics within the target region, such as the mean[8], follow the same general form in (8). Therefore, a similar approach could be used in cases where we have prior knowledge on such image statistics.

Our direct penalty-based approach for inequality constraints yields a considerable increase in performance with respect to to Lagrangian-dual optimization [1], while being faster and more stable. We hypothesize that this is due, in part, to the *interplay* between stochastic optimization (e.g., stochastic gradient descent) for the primal and the iterates/projections for the Lagrangian dual[9]. Such dual iterates/projections are basic (non-stochastic) gradient methods for handling the constraints. Basic gradient methods have well-known issues with deep networks, e.g. they are sensitive to the learning rate and prone to weak local minima. Therefore, the dual part in Lagrangian optimization might obstruct the practical and theoretical benefits of stochastic optimization (e.g., speed and strong generalization performance), which are widely established for unconstrained deep network losses [50]. Our penalty-based approach transforms a constrained problem into an unconstrained loss, thereby handling the constraints fully within stochastic optimization and avoiding completely the dual steps. While penalty-based approaches do not guarantee constraint satisfaction, our work showed that they can be extremely useful in the context of constrained CNN segmentation.

[8] Notice that the mean of intensity within the target region can be represented with network output using general form (8), with f_p corresponding to the intensity of pixel p

[9] In fact, a similar hypothesis was made in [25] to explain the negative results of Lagrangian optimization in the case of equality constraints.

7. Conclusion

In this paper, a novel loss function is present for weakly supervised image segmentation, which, despite its simplicity, performs significantly better than Lagrangian optimization for this task. We achieve results close to full supervision by annotating only a small fraction of the pixels, across three different tasks, and with negligible computation overhead. While our experiments focused on basic linear constraints such as the target-region size and image tags, our direct constrained-CNN loss can be easily extended to other non-linear constraints, e.g., invariant shape moments [2] or other region statistics [3]. Therefore, it has the potential to close the gap between weakly and fully supervised learning in semantic medical image segmentation.

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.

References

References

1. D. Pathak, P. Krahenbuhl, T. Darrell, Constrained convolutional neural networks for weakly supervised segmentation, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1796–1804.

2. M. Klodt, D. Cremers, A convex framework for image segmentation with moment constraints, in: IEEE International Conference on Computer Vision (ICCV), 2011, pp. 2236–2243.

3. Y. Lim, K. Jung, P. Kohli, Efficient energy minimization for enforcing label statistics, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (9) (2014) 1893–1899.

4. J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

5. J. Dolz, C. Desrosiers, I. Ben Ayed, 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study, NeuroImage 170 (2018) 456–470.

6. G. J. S. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical Image Analysis 42 (2017) 60–88.

7. J. Dai, K. He, J. Sun, Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1635–1643.

8. A. L. Bearman, O. Russakovsky, V. Ferrari, F. Li, What's the point: Semantic segmentation with point supervision, in: European Conference on Computer Vision (ECCV), 2016, pp. 549–565.

9. D. Lin, J. Dai, J. Jia, K. He, J. Sun, Scribblesup: Scribble-supervised convolutional networks for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 3159–3167.

10. M. Tang, A. Djelouah, F. Perazzi, Y. Boykov, C. Schroers, Normalized Cut Loss for Weakly-supervised CNN Segmentation, in: IEEE conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, 2018.

11. P. O. Pinheiro, R. Collobert, From image-level to pixel-level labeling with convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1713–1721.

12. Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, S. Yan, Stc: A simple to complex framework for weakly-supervised semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (11) (2017) 2314–2320.

13. J. Weston, F. Ratle, H. Mobahi, R. Collobert, Deep learning via semi-supervised embedding, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 639–655.

14. I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT press, 2016.

15. M. Tang, F. Perazzi, A. Djelouah, I. B. Ayed, C. Schroers, Y. Boykov, On regularized losses for weakly-supervised cnn segmentation, in: European Conference on Computer Vision (ECCV), 2018.

16. M. Rajchl, M. C. Lee, O. Oktay, K. Kamnitsas, J. Passerat-Palmbach, W. Bai, M. Damodaram, M. A. Rutherford, J. V. Hajnal, B. Kainz, et al., Deepcut: Object segmentation from bounding box annotations using convolutional neural networks, IEEE Transactions on Medical Imaging 36 (2) (2017) 674–683.

17. G. Papandreou, L.-C. Chen, K. P. Murphy, A. L. Yuille, Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1742–1750.

18. A. Kolesnikov, C. H. Lampert, Seed, expand and constrain: Three principles for weakly-supervised image segmentation, in: European Conference on Computer Vision, Springer, 2016, pp. 695–711.

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

20. M. Niethammer, C. Zach, Segmentation with area constraints, Medical Image Analysis 17 (1) (2013) 101–112.

21. L. Gorelick, F. R. Schmidt, Y. Boykov, Fast trust region for segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 1714–1721.

22. Y. Zhang, P. David, B. Gong, Curriculum domain adaptation for semantic segmentation of urban scenes, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2039–2049.

23. Y. Boykov, H. N. Isack, C. Olsson, I. B. Ayed, Volumetric bias in segmentation and reconstruction: Secrets and solutions, in: IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1769–1777.

24. Z. Jia, X. Huang, I. Eric, C. Chang, Y. Xu, Constrained deep weak supervision for histopathology image segmentation, IEEE Transactions on Medical Imaging 36 (11) (2017) 2376–2388.

25. P. Márquez-Neila, M. Salzmann, P. Fua, Imposing hard constraints on deep networks: Promises and limitations, arXiv preprint arXiv:1706.02025.

26. S. N. Ravi, T. Dinh, V. S. R. Lokhande, V. Singh, Constrained deep learning using conditional gradient and applications in computer vision, arXiv preprint arXiv:1803.06453.

27. S. Zhang, A. Constantinides, Lagrange programming neural networks, IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing 39 (7) (1992) 441–452.

28. J. C. Platt, A. H. Barr, Constrained differential optimization, in: Neural Information Processing Systems, 1988, pp. 612–621.

29. A. Vezhnevets, V. Ferrari, J. M. Buhmann, Weakly supervised semantic segmentation with a multi-image model, in: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE, 2011, pp. 643–650.

30. J. M. Buhmann, V. Ferrari, A. Vezhnevets, Weakly supervised structured output learning for semantic segmentation, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2012, pp. 845–852.

31. J. Verbeek, B. Triggs, Region classification with markov field aspect models, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2007, pp. 1–8.

32. A. Vezhnevets, J. M. Buhmann, Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 3249–3256.

33. C. Rother, V. Kolmogorov, A. Blake, Grabcut: Interactive foreground extraction using iterated graph cuts, in: ACM transactions on graphics (TOG), Vol. 23, ACM, 2004, pp. 309–314.

34. D. Pathak, E. Shelhamer, J. Long, T. Darrell, Fully convolutional multi-class multiple instance learning, in: ICLR Workshop, 2015.

35. J. Xu, A. G. Schwing, R. Urtasun, Tell me what you see and i will show you where it is, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 3190–3197.

36. A. Khoreva, R. Benenson, J. H. Hosang, M. Hein, B. Schiele, Simple does it: Weakly supervised instance and semantic segmentation., in: CVPR, Vol. 1, 2017, p. 3.

37. J. Xu, A. G. Schwing, R. Urtasun, Learning to segment under various forms of weak supervision, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3781–3790.

38. P. Vernaza, M. Chandraker, Learning random-walk label propagation for weakly-supervised semantic segmentation, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Vol. 3, 2017, p. 3.

39. A. Arnab, S. Zheng, S. Jayasumana, B. Romera-Paredes, M. Larsson, A. Kirillov, B. Savchynskyy, C. Rother, F. Kahl, P. H. Torr, Conditional random fields meet deep neural networks for semantic segmentation: Combining probabilistic graphical models with deep learning for structured prediction, IEEE Signal Processing Magazine 35 (1) (2018) 37–52.

40. L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Semantic image segmentation with deep convolutional nets and fully connected crfs, in: ICLR, 2015.

41. O. Chapelle, B. Schölkopf, A. Zien, Semi-Supervised Learning (Adaptive Computation and Machine Learning Series), The MIT Press, 2006.

42. M. Rajchl, M. C. Lee, F. Schrans, A. Davidson, J. Passerat-Palmbach, G. Tarroni, A. Alansary, O. Oktay, B. Kainz, D. Rueckert, Learning under distributed weak supervision, arXiv preprint arXiv:1606.01100.

43. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, et al., Slic superpixels compared to state-of-the-art superpixel methods, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (11) (2012) 2274–2282.

44. A. Paszke, A. Chaurasia, S. Kim, E. Culurciello, Enet: A deep neural network architecture for real-time semantic segmentation, arXiv preprint arXiv:1606.02147.

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

46. T. M. Quan, D. G. Hildebrand, W.-K. Jeong, Fusionnet: A deep fully residual convolutional neural network for image segmentation in connectomics, arXiv preprint arXiv:1612.05360.

47. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in pytorch.

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

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

50. M. Hardt, B. Recht, Y. Singer, Train faster, generalize better: Stability of stochastic gradient descent, in: International Conference on Machine Learning (ICML), 2016, pp. 1225–1234.