Reducing the Hausdorff Distance in Medical Image Segmentation with Convolutional Neural Networks

Davood Karimi, and Septimiu E. Salcudean, *Fellow, IEEE*

***Abstract*—The Hausdorff Distance (HD) is widely used in evaluating medical image segmentation methods. However, existing segmentation methods do not attempt to reduce HD directly. In this paper, we present novel loss functions for training convolutional neural network (CNN)-based segmentation methods with the goal of reducing HD directly. We propose three methods to estimate HD from the segmentation probability map produced by a CNN. One method makes use of the distance transform of the segmentation boundary. Another method is based on applying morphological erosion on the difference between the true and estimated segmentation maps. The third method works by applying circular/spherical convolution kernels of different radii on the segmentation probability maps. Based on these three methods for estimating HD, we suggest three loss functions that can be used for training to reduce HD. We use these loss functions to train CNNs for segmentation of the prostate, liver, and pancreas in ultrasound, magnetic resonance, and computed tomography images and compare the results with commonly-used loss functions. Our results show that the proposed loss functions can lead to approximately $18-45\%$ reduction in HD without degrading other segmentation performance criteria such as the Dice similarity coefficient. The proposed loss functions can be used for training medical image segmentation methods in order to reduce the large segmentation errors.**

***Index Terms*—Hausdorff distance, loss functions, medical image segmentation, convolutional neural networks**

I. Introduction

IMAGE segmentation is the process of delineating an object or region of interest in an image. It is a central task in medical image analysis, where the volume of interest has to be isolated for visualization or further analysis. Some of the applications of medical image segmentation include measuring the size or shape of the volume of interest, creating image atlases, targeted treatment, and image-guided intervention.

Medical image segmentation has been the subject of numerous papers in recent decades. In many applications, the manual segmentation produced by an expert radiologist is still regarded as the gold standard. However, compared with manual segmentation, computerized semi-automatic and fully-automatic segmentation methods have the potential for increasing the speed and reproducibility of the results [1], [2]. Fully-automatic segmentation methods eliminate the inter-observer and intra-observer variability that are caused by such factors as the differences in expertise and attention and errors due to visual fatigue. Moreover, especially with the emergence of convolutional neural network (CNN)-based segmentation algorithms in recent years, great progress has been made in reducing the performance gap between automatic and manual segmentation methods [3], [4].

The performance of automatic segmentation methods is usually evaluated by computing some common objective criteria such as the Dice similarity coefficient (DSC), mean boundary distance, volume difference or overlap, and Hausdorff Distance (HD) [5], [6]. Among these, HD is one of the most informative and useful criteria because it is an indicator of the largest segmentation error. In some applications, segmentation is one step in a more complicated multi-step process. For example, some multimodal medical image registration methods rely on segmentation of an organ of interest in one or several images. In such applications, the largest segmentation error as quantified by HD can be a good measure of the usefulness of the segmentations for the intended task. As illustrated in Figure 1, for two point sets X and Y, the one-sided HD from X to Y is defined as [7]:

$$\mathrm{hd}(X,Y) = \max_{x\in X} \min_{y\in Y} \|x-y\|_2, \tag{1}$$

and similarly for $\mathrm{hd}(Y,X)$:

$$\mathrm{hd}(Y,X) = \max_{y\in Y} \min_{x\in X} \|x-y\|_2. \tag{2}$$

The bidirectional HD between these two sets is then:

$$\mathrm{HD}(X,Y) = \max\big(\mathrm{hd}(X,Y), \mathrm{hd}(Y,X)\big) \tag{3}$$

In the above definitions we have used the Euclidean distance, but other metrics can be used instead. Intuitively, $\mathrm{HD}(X,Y)$ is the longest distance one has to travel from a point in one of the two sets to its closest point in the other set. In image segmentation, HD is computed between boundaries of the estimated and ground-truth segmentations, which consist of curves in 2D and surfaces in 3D.

Although HD is used extensively in evaluating the segmentation performance, segmentation algorithms rarely aim at minimizing or reducing HD directly [8], [9]. For example in the segmentation methods based on deformable models, the typical formulation of the external energy used to drive the segmentation algorithm is an integral (i.e., sum) of the edge information along the segmentation boundary [10], [11]. Therefore, these methods can be interpreted as minimizing the mean error over the segmentation boundary. Atlas-based segmentation methods, which are another class of widely-used techniques, work by registering a set of reference images to

D. Karimi is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, V6T 1Z4, Canada, e-mail: karimi@ece.ubc.ca

S. E. Salcudean is with the Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada.

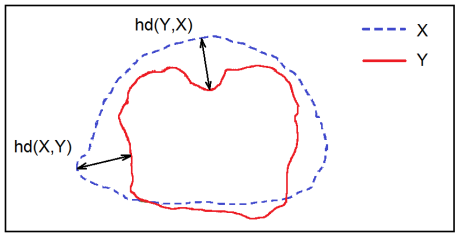


Figure 1. A schematic showing the Hausdorff Distance between points sets X and Y.

the target image by minimizing such global loss functions as the sum of squared difference of image intensities or the mutual information [12], [13]. Similarly, machine learning-based image segmentation methods aim at reducing a global loss function rather than the largest segmentation error [14], [15].

To the best of our knowledge, with one exception [16], no previous study has proposed a method for directly minimizing or reducing HD in medical image segmentation. There may be several reasons why previous works have not targeted HD. One reason is that unlike many other criteria such as cross-entropy, DSC, and volume overlap that are affected by the segmentation performance over the entire image, HD is determined solely by the largest error. If a segmentation algorithm is designed to focus on the largest error, the overall segmentation performance may suffer. Moreover, an algorithm that aims solely at minimizing the largest error may be unstable. This is confirmed by our own observations in this study, which are discussed in Section III of this paper. Moreover, especially for segmentation of complex structures that are common in medical imaging, a segmentation algorithm may achieve satisfying accuracy over most of the image but have large errors at one or a few isolated locations. This could occur because of different reasons such as weak or missing edges, artifacts, or low signal to noise ratio. In these cases, accurate segmentation is difficult or impossible even for a human expert. Hence, it may not be reasonable to expect high segmentation accuracy everywhere in the image because "the ground truth" may be unreliable or nonexistent. The sensitivity of HD to noise and outliers has been well documented in the computer vision literature. For image matching, for example, it has been suggested to use modified definitions of HD or to combine HD with other image information to obtain more robust algorithms [17], [18]. Two widely-used variations of HD that have been designed to reduce the sensitivity to outliers are Partial HD [19] and Modified HD [17]. Partial HD replaces the $\max$ operation in Equation (1) with the K^{th} largest value, whereas Modified HD replaces it with the averaging operation.

Moreover, direct minimization of HD is very challenging from an optimization viewpoint. Most of the studies in computer vision that have used HD have focused on a restricted set of problems such as object matching or face detection. In these applications, the goal is to match a template B to an image A subject to simple transformations such as translation, rotation, and scaling. Hence, the goal is to find a small set of parameters p such that the HD between the transformed B and A, i.e., HD$(T_p(B), A)$, is minimized. Even this restricted scenario is not easy to handle. Some studies have suggested methods such as genetic algorithms [20], while others have used exhaustive search to solve the problem [21], [22]. A number of studies have proposed similar formulations for medical image registration by approximately minimizing the HD between reference and target landmarks under rigid transformations [23], [24]. We are aware of only one work that has used HD in medical image segmentation [16]. That study was quite different from the methods proposed in this work. In particular, the authors of [16] focus on the specific problem of multi-surface segmentation where each small surface is nested within a larger surface. Moreover, their segmentation method is based on minimizing an energy function that consists of a data term and a smoothness term. Instead of minimizing HD, they propose to use the prior knowledge about the maximum value of HD as a constraint. They show that the resulting problem could be NP-hard and suggest simplifying assumptions in order to obtain an approximate solution. They apply their method for segmentation of different structures in MR and ultrasound images. However, they only visually display their results and do not provide any quantitative evaluation. They also do not compare their method against any other methods.

The goal of this paper is to propose methods for reducing HD in CNN-based segmentation methods. CNN-based methods are relative new-comers to the field of medical image segmentation, but they have already proved to be highly versatile and effective [25], [3]. These methods usually produce a dense (i.e., pixel- or voxel-wise) segmentation probability map of the organ or volume of interest, although there are some exceptions [26], [27]. The early CNN-based image segmentation methods applied a soft-max function to the output layer activations and defined a loss function in terms of the negative log-likelihood, which is equivalent to a cross-entropy loss for binary segmentation [28]. Later, some studies proposed different loss functions to address specific challenges of medical image segmentation. For example, some works suggested a weighted cross-entropy, where larger weights are assigned to more important regions such as the boundaries of the volume of interest [29], [30]. Another difficulty in medical image segmentation is that often the object of interest occupies a small portion of the image, biasing the algorithm towards achieving higher specificity than sensitivity. To counter this effect, it was suggested that DSC be used as the objective function [31]. Another study has suggested that more control over sensitivity and specificity can be achieved by using the Tversky Index as the objective function [32]. These studies have shown that the choice of the loss function can have a large impact on the performance of image segmentation methods. Recently, some studies have argued that the choice of a good loss function for training of deep learning models has been unfairly neglected and that research on this topic can lead to large improvements in the performance of these models [33].

In this paper, we propose techniques for reducing HD in CNN-based medical image segmentation. The novel aspects of this work are as follows: 1) We propose three different loss functions based on HD that, to the best of our knowledge, are novel and have not been used for medical image segmentation

before, 2) We use four datasets to segment different organs in different medical imaging modalities and empirically show that using these loss functions can significantly reduce large segmentation errors, 3) Through extensive experiments we show the potential benefits and challenges of using HD-based loss functions for medical image segmentation.

The paper is organized as follows. In Section II, we propose three methods for estimating HD from the output probability map of a CNN. Because minimizing HD directly may not be desirable and could lead to unstable training, based on each of the three methods of estimating HD we propose an "HD-inspired" loss function that can be used for stable training. After explaining our methods in Section II, we will present and discuss our results on four different medical image datasets in Section III. We will describe the conclusions of this work in Section IV.

II. Materials and Methods

A. Notations

We denote the output segmentation probability map of a CNN with $q \in [0, 1]$. To obtain a binary segmentation of the object versus the background, one usually thresholds q at 0.50 to get a segmentation map with values in $\{0, 1\}$, where 0 indicates the background and 1 indicates the object. We denote this binary map with $\bar{q}$. Similarly, we denote the ground-truth segmentation with $p \in [0, 1]$ and $\bar{p} \in \{0, 1\}$, although typically the ground-truth segmentation is a binary map, i.e., $p \equiv \bar{p}$. As shown in Figure 2, we denote the boundaries of $\bar{p}$ and $\bar{q}$ with δp and δq, respectively. For ease of illustration, in this section we use 2D figures to explain our proposed methods. However, the extension of the methods to 3D is trivial, and we will present experimental results with 2D as well as 3D images in Section III.

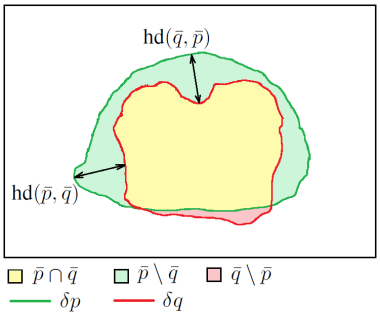
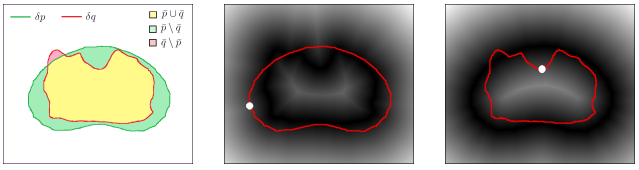


Figure 2. A visual depiction of some of the notations used in this paper.

B. Estimation of the Hausdorff Distance Based on Distance Transforms

Our first approximation of HD is based on distance transforms (DT). The DT of a digital image is a derived representation of that image where each pixel has a value equal to its distance to an object of interest in the image. For a 2D binary image $X[i, j]$, with 0 representing the background and 1 indicating the object, we have [34]:



(a) (b) (c)

Figure 3. (a) An example of a 2D ground-truth and predicted segmentations, denoted with $\bar{p}$ and $\bar{q}$, respectively. (b) The distance transform d_q with δp overlaid in red. The white circle shows the location of the largest d_q, which corresponds to hd$(\delta p, \delta q)$. (c) Similar to (b), for finding hd$(\delta q, \delta p)$.

$$\mathrm{DT}_X[i, j] = \min_{k,l;X[k,l]=1} d\big([i, j], [k, l]\big) \tag{4}$$

where d denotes the distance between pixel locations and $[k, l]$ denote the indices of the object (i.e., foreground) pixels.. In this work, we use the standard choice of the Euclidean distance: $d\big([i, j], [k, l]\big) = \sqrt{(k - i)^2 + (l - j)^2}$.

Here, we define the distance map of the ground-truth segmentation as the unsigned distance to the boundary, δp, and denote it with d_p. Similarly, d_q is defined as the distance to δq. As shown in Figure 3, it is clear that we can write:

$$\mathrm{hd}_{\mathrm{DT}}(\delta q, \delta p) = \max_{\Omega} \big((\bar{p} \triangle \bar{q}) \circ d_p\big) \tag{5}$$

where we have used the subscript DT to indicate that HD is computed using the distance transforms. In the above equation, and in the rest of this paper, $\triangle$ denotes the set operation of symmetric difference defined as $\bar{p} \triangle \bar{q} = (\bar{p} \setminus \bar{q}) \cup (\bar{q} \setminus \bar{p})$ [35]. For us, this can be simply computed as $\bar{p} \triangle \bar{q} = | \bar{p} - \bar{q} |$. Moreover, in the above equation, $\circ$ denotes the Hadamard (i.e., entry-wise) product and Ω denotes the grid on which the image is defined, which means that max is with respect to all pixels.

We can similarly compute $\mathrm{hd}_{\mathrm{DT}}(\delta p, \delta q)$, and then $\mathrm{HD}_{\mathrm{DT}}(\delta q, \delta p)$ as follows:

$$\mathrm{hd}_{\mathrm{DT}}(\delta p, \delta q) = \max_{\Omega} \big((\bar{p} \triangle \bar{q}) \circ d_q\big) \tag{6}$$

$$\mathrm{HD}_{\mathrm{DT}}(\delta q, \delta p) = \max\big(\mathrm{hd}_{\mathrm{DT}}(\delta q, \delta p), \mathrm{hd}_{\mathrm{DT}}(\delta p, \delta q)\big) \tag{7}$$

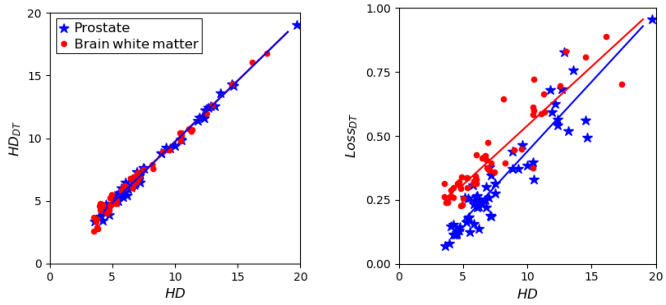
Figure 4(a) illustrates that this method is a correct way of computing HD. In this figure, we have plotted the HD estimated using Equation (7) versus the exact HD computed using [36] between the ground-truth and approximate segmentations of 50 3D Magnetic Resonance (MR) prostate images and 50 brain white matter MR images. For both prostate and brain data the Pearson correlation coefficient of the fitted linear function is above 0.99. Based on the above estimation of HD, we suggest the following loss function for CNN training:

$$\mathrm{Loss}_{\mathrm{DT}}(q, p) = \frac{1}{|\,\Omega\,|} \sum_{\Omega} \Big((p - q)^2 \circ (d_p^{\alpha} + d_q^{\alpha})\Big) \tag{8}$$

Compared with Equation (7) that is an accurate estimator of HD, this loss function is different in three aspects. First, instead of focusing only on the largest segmentation error, we smoothly penalize larger segmentation errors. The parameter

α determines how strongly we penalize larger errors. To determine a good value for this parameter, we tried values of $\alpha \in \{0.5, 1.0, ..., 3.5, 4.0\}$ in small cross-validation experiments. Our experiments showed that values of α between 1.0 and 3.0 led to good results. In all the experiments reported for this method in this paper, we used a value of $\alpha = 2.0$, which we found to be the best value in that set. Second, unlike Equation (7), we use p and q instead of the thresholded maps, $\bar{p}$ and $\bar{q}$, to allow the training to take into account this useful information. Finally, instead of $\mid p-q \mid$, we use $(p-q)^2$. This choice is inspired by the results of [33] and will be justified empirically in Section III.

Figure 4(b) shows a plot of Loss_{DT} versus exact HD for the same prostate and brain MR data as in 4(a). The loss functions have been scaled to $[0, 1]$ for display. For both prostate and brain data the Pearson correlation coefficient for the fitted linear function is approximately 0.93. It is clear that there is a strong correlation between the two, so that reducing Loss_{DT} should lead to a decrease in HD.



(a) (b)

Figure 4. Plots of HD_{DT} and Loss_{DT} versus exact HD for sets of 3D MR prostate and brain white matter images and their rough segmentations.

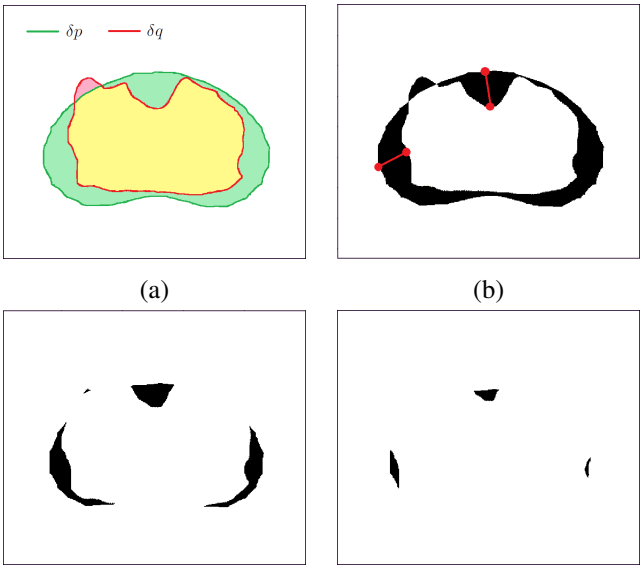
A drawback of this method is the high computational cost of computing the distance transforms, d_p and d_q. In this work, we used the algorithm proposed in [37] for computing DT in 2D and the algorithm in [38] for experiments with 3D images. One may use less accurate but faster algorithms instead, because very accurate estimation of DT is not needed in this application. Nonetheless, the computational cost will remain high, especially in 3D. Moreover, the cost will be much higher for computing d_q than for d_p. This is because q changes during training and therefore d_q should be recomputed for all images after each training epoch. On the other hand, d_p needs to be computed only once. Therefore, one way of reducing the computational cost is to only consider the one-sided HD, $\text{hd}_{\text{DT}}(\delta q, \delta p)$. This leads to the following modified loss function (where we use OS to indicate "one-sided"):

$$\text{Loss}_{\text{DT-OS}}(q,p) = \frac{1}{\mid \Omega \mid} \sum_{\Omega} \left((p-q)^2 \circ d_p^{\alpha} \right) \tag{9}$$

We will present some experimental results with this loss function as well in Section III.

C. Estimation of the Hausdorff Distance Using Morphological Operations

Although the distance transform-based method explained above is simple and intuitive, it has a high computational cost. In this section, we propose an alternative approach that is based on the use of morphological operations. As can be seen in Figure 5, $\text{HD}(\delta q, \delta p)$ is roughly related to the largest thickness of the difference between the true and estimated segmentations, $\bar{p} \triangle \bar{q}$. Therefore, one can obtain an approximate estimation of $\text{HD}(\delta q, \delta p)$ by applying morphological erosion on $\bar{p} \triangle \bar{q}$.



(a) (b)

(c) (d)

Figure 5. (a) An example of a 2D ground-truth and predicted segmentations, denoted with $\bar{p}$ and $\bar{q}$, respectively. (b) The map of $\bar{p} \triangle \bar{q}$ for this example; $\text{hd}(\delta p, \delta q)$ and $\text{hd}(\delta q, \delta p)$ have been marked with red line segments on this figure. (c) and (d) The eroded map of $\bar{p} \triangle \bar{q}$ after applying, respectively, 5 and 10 erosions with a cross-shaped structuring element of size 5.

Morphological erosion of a binary object S defined on a grid Ω using a structuring element B is defined as [34]:

$$S \ominus B = \{z \in \Omega | B(z) \subseteq S\} \tag{10}$$

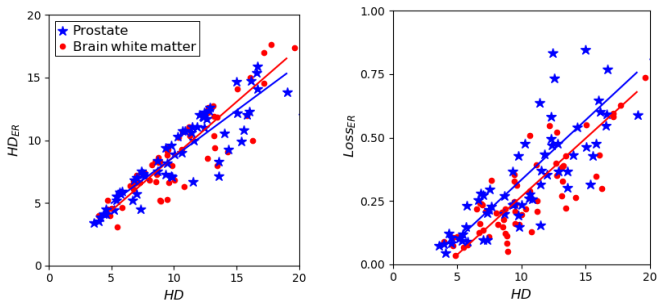
where $B(z)$ is the structuring element shifted on the grid such that it is centered on z.

Let us denote a structuring element with radius r as B_r. We suggest the following approximation to $\text{HD}(\delta q, \delta p)$ based on a morphological erosion of $\bar{p} \triangle \bar{q}$.

$$\begin{aligned} \text{HD}_{\text{ER}}(\delta q, \delta p) &= 2r^* \\ \text{where } & r^* = \text{minimum } r \\ \text{such that } & (\bar{p} \triangle \bar{q}) \ominus B_r = \emptyset \end{aligned} \tag{11}$$

where the subscript ER indicates that the Hausdorff Distance is computed using morphological erosion.

HD_{ER} defined above is a lower bound on the true HD because if erosion of $\bar{p} \triangle \bar{q}$ with B_{r^*} does not result in an empty set then $\text{HD} > r^*$. One can make up pathological examples for which HD is much larger than HD_{ER}. However, as can be seen in Figure 6(a), in practice the proposed HD_{ER} is a



(a) (b)

Figure 6. Plots of HD_{ER} and Loss_{ER} versus exact HD for sets of 3D MR prostate and brain white matter images and their rough segmentations.

good approximation of the exact HD. The Pearson correlation coefficient for the fitted linear function in this figure for the prostate and brain data is 0.93 and 0.91, respectively.

As can be seen by comparing Figures 4(a) and 6(a), HD_{ER} is not as accurate as HD_{DT}. However, an advantage of HD_{ER} is that morphological operations can be implemented efficiently using convolutional operations and thresholding. This had been demonstrated long before the recent surge of interest in CNNs [39], [40]. Therefore, HD_{ER} can be computed more efficiently than HD_{DT}. Similar to what we did for HD_{DT} above, instead of using HD_{ER} directly as the loss function, we propose the following relaxed loss function:

$$\text{Loss}_{\text{ER}}(q,p) = \frac{1}{|\,\Omega\,|}\sum_{k=1}^{K}\sum_{\Omega}\big((p-q)^2 \ominus_k B\big)k^{\alpha} \qquad (12)$$

In the above equation we have used $\ominus_k$ to denote k successive erosions. Note that erosion is applied to $(p-q)^2$, which is not binary. This is based on the generalized definition of erosion proposed in [39]. In our work, we compute this via convolution with a kernel whose elements sum to one followed by a soft thresholding [41] at 0.50. For 2D, we use a cross-shaped structuring element $B = \begin{pmatrix} 0 & 1/5 & 0 \\ 1/5 & 1/5 & 1/5 \\ 0 & 1/5 & 0 \end{pmatrix}$. Similarly, for 3D we use a convolutional kernel of size 3 with the center element and its 6-neighbors set to $1/7$ and the remaining 20 elements set to zero. In Equation (12), K denotes the total number of erosions. Increasing K will increase the computational cost. On the other hand, K should be large enough because all parts of $\bar{p} \triangle \bar{q}$ that remain after K erosions will be weighted equally. In practice, one has to set K based on the expected range of segmentation errors. We set $K = 10$ for all experiments in this work. The parameter α determines how strongly we penalize larger segmentation errors. We used $\alpha = 2.0$ in our experiments. Similar to the method in Section II-B, we chose this value using a grid search.

Figure 6(b) shows a plot of Loss_{ER} versus the exact HD on a set of images. There is a good correlation between Loss_{ER} and HD. The Pearson correlation coefficient for the fitted linear function in this figure for both prostate and brain data is 0.83. One can easily compute Loss_{ER} by stacking K convolutional layers to the end of any CNN.

D. Estimation of Hausdorff Distance using Convolutions with Circular/Spherical Kernels

Let us denote a circular-shaped convolutional kernel of radius r with B_r. Elements of B_r are normalized such that they sum to one. Then we can write:

$$\begin{aligned} &\text{hd}_{\text{CV}}(\delta q, \delta p) = \max(r_1, r_2) \\ &\quad \text{where} \;\; r_1 = \max r \\ &\qquad \text{such that} \;\; \max_{\Omega} f_h(\bar{p}^C * B_r) \circ (\bar{q} \setminus \bar{p}) > 0 \\ &\quad \text{and} \;\; r_2 = \max r \\ &\qquad \text{such that} \;\; \max_{\Omega} f_h(\bar{p} * B_r) \circ (\bar{p} \setminus \bar{q}) > 0 \end{aligned} \qquad (13)$$

where we have used the subscript CV to denote the Hausdorff Distance computed using convolutions. In the above equation, $\bar{p}^C = 1 - \bar{p}$ denotes the complement of $\bar{p}$, and f_h is a hard thresholding function that sets all values below 1 to zero. The schematic in Figure 7 helps the reader understand this equation. It shows that HD can be computed using only convolution and thresholding operations. We can compute $\text{hd}_{\text{CV}}(\delta p, \delta q)$ using a similar equation and then $\text{HD}_{\text{CV}}(\delta p, \delta q) = \max\big(\text{hd}_{\text{CV}}(\delta q, \delta p), \text{hd}_{\text{CV}}(\delta p, \delta q)\big)$.

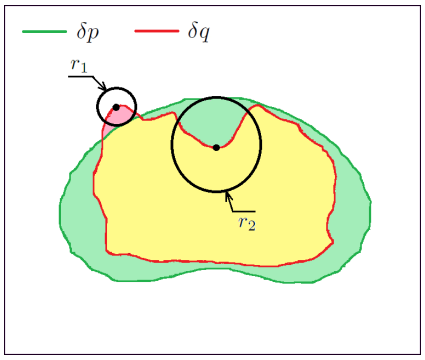


Figure 7. A schematic illustration of the method to compute HD using convolutions with circular kernels. Circles with radii r_1 and r_2 show the convolutional kernels that determine HD according to Equation (13). From this figure and Equation (13) one can see that $\text{hd}_{\text{CV}}(\delta q, \delta p) = \max(r_1, r_2)$.

As shown in Figure 8(a), HD_{CV} is an accurate approximation of the true HD. The Pearson correlation coefficient for the fitted linear function in this figure for both prostate and brain data is approximately 0.99.

We should note that Equation (13) provides an exact estimate of HD in the continuous domain. However, on a discrete grid, the precision of HD computation is limited by the pixel size. Moreover, there is some discretization error when representing circular/spherical convolutional kernels on a discrete grid. This error is larger for smaller circles/spheres. As a result, the spread from the straight line is greater for smaller HD values in Figure 8(a).

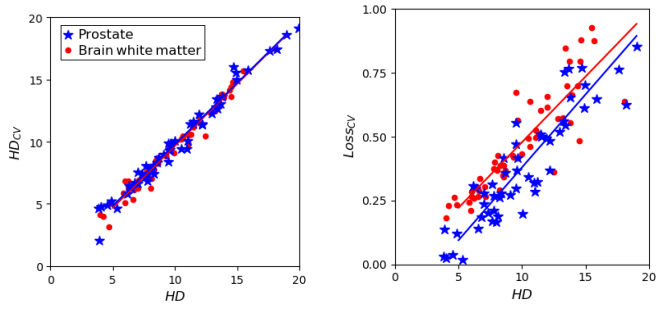
Once again, we aim for a relaxed loss function that smoothly penalizes larger errors instead of focusing only on the largest error. Therefore, we suggest the following loss function:

$$\text{Loss}_{\text{CV}}(q,p) = \frac{1}{|\,\Omega\,|}\sum_{r\in R} r^{\alpha}\sum_{\Omega}\big[f_s(B_r * \bar{p}^C)\circ f_{\bar{q}\setminus\bar{p}} + f_s(B_r * \bar{p})\circ f_{\bar{p}\setminus\bar{q}} + f_s(B_r * \bar{q}^C)\circ f_{\bar{p}\setminus\bar{q}} + f_s(B_r * \bar{q})\circ f_{\bar{q}\setminus\bar{p}}\big] \quad (14)$$

where $f_{\bar{q}\setminus\bar{p}}$ is a relaxed estimation of $\bar{q}\setminus\bar{p}$ defined as:

$$f_{\bar{q}\setminus\bar{p}} = (p-q)^2 q \quad (15)$$

and similarly for $f_{\bar{p}\setminus\bar{q}}$. Moreover, we have replaced the hard thresholding, f_h, in Equation (13) with the soft thresholding f_s in Equation (14). The parameter α plays the same role here as it did in Equations (9) and (12). Similar to the above two methods, we chose $\alpha = 2.0$ via a grid search in the range $[0.50, 4.0]$. Figure 8(b) shows Loss_{CV} as a function of the exact HD. The Pearson correlation coefficient for the fitted linear function in this figure for the prostate and brain data is approximately 0.91 and 0.88, respectively.



(a) (b)

Figure 8. Plots of HD_{CV} and Loss_{CV} versus exact HD for sets of 3D MR prostate and brain white matter images and their rough segmentations.

Similar to Loss_{ER}, Loss_{CV} is based on convolution and thresholding operations, which can be implemented easily in deep learning software frameworks. A comparison of Figures 6(a) and 8(a) shows that HD_{CV} is a more accurate estimation of HD than HD_{ER} is. However, whereas HD_{ER} is computed using small fixed convolutional kernels (of size 3, in our implementation), computation of Loss_{CV} will require applying filters of increasing size. For very large convolutional filters, especially in 3D, the computational load can become very significant. Therefore, in computing Loss_{CV} we use a maximum kernel radius of 18 pixels in 2D and 9 voxels in 3D. Hence, larger segmentation errors are treated equally. To further reduce the cost, we do not use kernels of every size, but only at steps of 3, because there is no need for such fine resolution. Therefore, in Equation (14) we set $R = \{3, 6, \ldots 18\}$ in our experiments with 2D images and $R = \{3, 6, 9\}$ in experiments with 3D images. In practice, one can choose R based on the expected range of segmentation errors and, of course, the pixel/voxel size.

E. Data

We used four datasets, one of 2D images and three of 3D images, in our experiments. A brief description of the data is provided below.

1) 2D ultrasound images of prostate: This dataset consisted of trans-rectal ultrasound (TRUS) images of 675 patients. From each patient, between 7 and 14 2D TRUS images of size 415×490 pixels with a pixel size of $0.15 \times 0.15\,\text{mm}^2$ had been acquired. The clinical target volume (CTV) had been delineated in each slice by experienced radiation oncologists using a semi-automatic segmentation software described in [42]. The use of this software biased the segmentation of the prostate at the base and apex. The "ground-truth" segmentation of the base and apex is also made unreliable due to the lack of clear landmarks. Therefore, we chose to work with the slices that belonged to the mid-gland, which we defined as the middle 40% of the prostate. As a result, we had a total of 1805 2D images from 450 patients for training and 820 images from the remaining 225 patients for test.

2) 3D MR images of prostate: A total of 80 training and 30 test images were included in this dataset. This included the training data from the PROMISE12 challenge [43] as well as the Medical Segmentation Decathlon challenge (https://decathlon.grand-challenge.org/). The liver and pancreas datasets described below were also obtained through the Medical Segmentation Decathlon challenge. The pre-processing applied on prostate MR images included bias correction [44], resampling to an isotropic voxel size of 1 mm^3, and cropping to a size of $128 \times 128 \times 96$ voxels.

3) 3D CT images of liver: This dataset consisted of 131 CT images. We used 100 images for training and 31 for test. As pre-processing, we linearly mapped the voxel values (in Hounsfield Units) from $[-1000, 1000]$ to $[0, 1]$, cropping voxels smaller than -1000 and larger than 1000. We then resampled the images to an isotropic voxel size of 2 mm^3, and cropped them to a size of $192 \times 192 \times 128$ voxels.

4) 3D CT images of pancreas: This dataset consisted of 282 CT images. We used 200 images for training and 82 for test. The pre-processing was similar to that for the liver CT images described above.

F. CNN Architecture and Training Procedures

Currently, the most common loss functions for CNN-based medical image segmentation are the cross-entropy and DSC. Our experience shows that, compared with cross-entropy, DSC consistently leads to better results. Therefore, we will compare the following loss functions in our experiments:

- DSC, defined as:

$$f_{\text{DSC}}(q,p) = 1 - \frac{2\sum_{\Omega}(p\circ q)}{\sum_{\Omega}(p^2+q^2)} \quad (16)$$

- Three different HD-based loss functions defined as:

$$f^{*}_{\text{HD}}(q,p) = \text{Loss}_{*}(q,p) + \lambda\left(1 - \frac{2\sum_{\Omega}(p\circ q)}{\sum_{\Omega}(p^2+q^2)}\right) \quad (17)$$

where $*$ is replaced with DT, ER, and CV to give three different *HD-based loss functions* $f_{\text{HD}}^{\text{DT}}(q,p)$, $f_{\text{HD}}^{\text{ER}}(q,p)$, and $f_{\text{HD}}^{\text{CV}}(q,p)$ based on the three losses in Equations (8), (12), and (14), respectively.

As can be seen in Equation (17), we augment our HD-based loss term with a DSC loss term. This results in a more stable training, especially at the start of the training. We choose λ such that equal weights are given to the HD-based and DSC loss terms. Specifically, after each training epoch, we compute the HD-based and DSC loss terms on the training data and set λ (for the next epoch) as the ratio of the mean of the HD-based loss term to the mean of the DSC loss term. This simple empirical approach ensured that both loss terms were given equal weight and it worked well in all of our experiments.

Because the goal of this work was to study the impact of the loss function, we decided to use a standard CNN architecture and training procedure. This allows us to reduce the impact of other factors that may confound the results. Hence, we used the U-net [29] and 3D U-net [45] for our experiments with 2D and 3D images, respectively. Data augmentation is very common in training deep learning models, especially when the training data is small. We used three standard data augmentation methods: 1) adding random noise to images, 2) random cropping, 3) random elastic deformations [31], [29]. On each of the four datasets, we used the same data augmentation parameters (e.g., the noise standard deviation and parameters of elastic deformation) for all loss functions.

All loss functions were minimized using the Adam optimizer [46]. We used the default parameter settings suggested in [46]. The only parameter that we tuned was the learning rate because its optimal value could depend on the loss function. To conduct a fair comparison of different loss functions, for evaluating each loss function on each dataset we performed the training for 10 learning rates logarithmically spaced in $[10^{-3}, 10^{-5}]$ for 50 epochs and chose the learning rate that achieved the lowest loss on the training data. The selected learning rate was typically in the range $[10^{-3}, 10^{-4}]$. After selecting the learning rate, the model was trained from scratch for a total of 100 epochs with the selected learning rate. The learning rate was divided by 2 whenever the training loss did not decrease by more than 1% in a training epoch.

All the models and training code were implemented in Python 3.6 and TensorFlow 1.2 and run in Linux. For the exhaustive search to select the learning rate we used an NVIDIA DGX1. For training the final models (upon choosing the learning rates) we used an NVIDIA GeForce GTX TITAN X so that the reported training times be relevant to the type of hardware that most researchers currently use.

III. Results and Discussion

Table I shows a summary of the results on the four datasets. We have used DSC and HD as the evaluation criteria. In addition to the mean and standard deviation of DSC and HD, we also report the 90th-percentile and maximum of HD on the test data. Moreover, average symmetric surface distance (ASD) values have been reported in the same table. In the last column of the same table, we have shown the training times. For DSC and HD, we performed paired t-tests on the test images to see if the results for different loss functions were significantly different. The results of these statistical tests have been shown using superscripts in this table; for each dataset, different superscripts indicate statistically significant difference at $p = 0.01$. As an example, for the 2D prostate ultrasound data these superscripts indicate that: 1) In terms of DSC, all loss functions are statistically similar, 2) In terms of HD, $f_{\text{HD}}^{\text{DT}}$ and $f_{\text{HD}}^{\text{CV}}$ are statistically similar and statistically different (lower HD) than both $f_{\text{HD}}^{\text{ER}}$ and f_{DSC}; moreover, $f_{\text{HD}}^{\text{ER}}$ is statistically different (lower HD) than f_{DSC}.

On all four datasets, the three HD-based loss functions have resulted in lower average HD on the test images. The reduction in the mean of HD ranges between 18% and 45%. In all cases, this reduction in HD is statistically significant. Also, with a few exceptions, the HD-based loss functions also reduced the maximum and 90th-percentile of HD on the test images. In many cases this reduction is as high as $30-50\%$. On the other hand, the paired t-tests did not show any significant differences in terms of DSC achieved by the proposed HD-based loss functions and the pure DSC loss. This summary clearly demonstrates that the proposed loss functions effectively reduce HD in CNN-based image segmentation.

Figure 9 shows example test images on which the proposed HD-based loss functions resulted in lower HD than the DSC loss. We have shown one example image from each dataset. For the 3D images, we have shown the axial slice on which the largest error occurred with the DSC loss function.

Based on the results in Table I, among the three HD-based loss functions, $f_{\text{HD}}^{\text{DT}}$ and $f_{\text{HD}}^{\text{CV}}$ resulted in lower HD than $f_{\text{HD}}^{\text{ER}}$. The difference was statistically significant on three out of the four datasets. On the other hand, the training times for $f_{\text{HD}}^{\text{DT}}$ and $f_{\text{HD}}^{\text{CV}}$ were longer than that for $f_{\text{HD}}^{\text{ER}}$. On average, $f_{\text{HD}}^{\text{DT}}$ led to the best results in terms of HD, but with training times on 3D images that were approximately twice those of $f_{\text{HD}}^{\text{ER}}$ and f_{DSC}. One may speculate that training with $f_{\text{HD}}^{\text{ER}}$ may lead to segmentation performance on par with $f_{\text{HD}}^{\text{DT}}$ and $f_{\text{HD}}^{\text{CV}}$ if it is given equal training time in hours, rather than equal number of training epochs. However, our experiments showed this was not the case. As shown in Figure 10, for all loss functions the segmentation performance on the test data plateaued well before 100 epochs of training.

There are other loss functions that could have been included in our experiments. For example, cross-entropy is also commonly used as a loss function for training deep learning-based image segmentation models. In our experiments, cross-entropy always performed worse than DSC. Using weighted cross-entropy did not significantly improve the results. For example, for the 2D prostate ultrasound data the DSC and HD achieved using weighted cross-entropy as the loss function were 0.919 ± 0.050 and 4.3 ± 3.2, respectively. For 3D CT pancreas data, the DSC and HD achieved using weighted cross-entropy as the loss function were 0.746 ± 0.125 and 34.0 ± 17.3, respectively.

We should note that although the training times for different loss functions are quite different, the test times are identical. This is because segmentation of a test image only requires a

Table I
SUMMARY OF THE RESULTS OF OUR EXPERIMENTS WITH FOUR DATASETS. FOR EACH DATASET, MEAN ± STANDARD DEVIATION OF DSC, HD, AND ASD ARE PRESENTED IN ADDITION TO THE MAXIMUM OF DSC, 90TH-PERCENTILE AND MAXIMUM OF HD, AND THE TRAINING TIME. SUPERSCRIPTS ON THE VALUES OF DSC AND HD INDICATE THE RESULTS OF PAIRED T-TESTS. FOR EACH DATASET, DIFFERENT SUPERSCRIPTS ON DSC AND HD INDICATE STATISTICALLY SIGNIFICANT DIFFERENCE AT $p = 0.01$.

Dataset	Loss Function	DSC	Maximum of DSC	HD (mm)	90th-percentile of HD (mm)	Maximum of HD (mm)	ASD	Training time (h)
2D prostate ultrasound	$f_{\text{DSC}}(q,p)$	0.932 ± 0.039 [a]	0.975	4.3 ± 2.8 [a]	7.5	14.4	1.52±0.90	**3.2**
	$f_{\text{HD}}^{\text{DT}}(q,p)$	$\mathbf{0.946 \pm 0.041}$ [a]	0.982	$\mathbf{2.6 \pm 1.8}$ [b]	**4.4**	**7.1**	$\mathbf{1.05 \pm 0.46}$	4.0
	$f_{\text{HD}}^{\text{ER}}(q,p)$	0.936 ± 0.041 [a]	**0.985**	3.0 ± 2.0 [c]	5.6	14.9	1.12±0.59	3.7
	$f_{\text{HD}}^{\text{CV}}(q,p)$	0.941 ± 0.036 [a]	0.977	2.7 ± 1.8 [b]	4.5	8.2	1.10±0.50	4.1
3D prostate MRI	$f_{\text{DSC}}(q,p)$	0.868 ± 0.046 [a]	0.925	7.5 ± 3.1 [a]	10.5	15.1	1.95±0.46	**8.0**
	$f_{\text{HD}}^{\text{DT}}(q,p)$	0.875 ± 0.042 [a]	**0.927**	$\mathbf{5.8 \pm 2.2}$ [b]	**7.6**	9.0	1.51±0.27	21
	$f_{\text{HD}}^{\text{ER}}(q,p)$	0.858 ± 0.046 [a]	0.921	6.1 ± 2.3 [b]	8.9	10.1	1.55±0.33	9.8
	$f_{\text{HD}}^{\text{CV}}(q,p)$	$\mathbf{0.876 \pm 0.040}$ [a]	0.923	$\mathbf{5.8 \pm 2.5}$ [b]	8.0	**8.4**	$\mathbf{1.49 \pm 0.27}$	14
3D Liver CT	$f_{\text{DSC}}(q,p)$	0.921 ± 0.048 [a]	0.949	46.8 ± 18.9 [a]	59.8	72.9	1.58±0.49	**22**
	$f_{\text{HD}}^{\text{DT}}(q,p)$	$\mathbf{0.940 \pm 0.040}$ [a]	0.959	$\mathbf{25.1 \pm 10.3}$ [b]	**38.4**	**41.2**	$\mathbf{1.37 \pm 0.39}$	47
	$f_{\text{HD}}^{\text{ER}}(q,p)$	0.936 ± 0.040 [a]	0.956	31.6 ± 12.1 [c]	44.4	50.2	1.44±0.41	26
	$f_{\text{HD}}^{\text{CV}}(q,p)$	0.935 ± 0.039 [a]	**0.966**	27.3 ± 13.4 [b]	41.8	43.8	1.38±0.41	42
3D Pancreas CT	$f_{\text{DSC}}(q,p)$	0.752 ± 0.120 [a]	0.855	32.1 ± 17.0 [a]	58.5	65.1	2.09±0.57	**22**
	$f_{\text{HD}}^{\text{DT}}(q,p)$	$\mathbf{0.784 \pm 0.059}$ [a]	**0.870**	$\mathbf{21.3 \pm 11.3}$ [b]	**35.2**	**37.7**	$\mathbf{1.84 \pm 0.44}$	50
	$f_{\text{HD}}^{\text{ER}}(q,p)$	0.767 ± 0.066 [a]	0.845	27.1 ± 13.6 [c]	41.6	45.0	1.98±0.43	24
	$f_{\text{HD}}^{\text{CV}}(q,p)$	0.780 ± 0.055 [a]	0.862	21.7 ± 11.0 [b]	39.0	44.1	1.91±0.39	34

forward pass through the network and does not involve the loss function in any way. The loss functions are only used during training. On an NVIDIA GeForce GTX TITAN X, the test time for a single image for the 2D prostate ultrasound, 3D prostate MRI, 3D liver CT, and 3D pancreas CT were, respectively 0.09, 0.36, 0.45, and 0.45 seconds. These times were identical for the networks trained with any of the HD-based loss functions as well as the DSC loss.

As we mentioned in Section II-B, with the distance transform-based approach, using a modified loss function based on the one-sided HD ($\text{Loss}_{\text{DT-OS}}$, shown in Equation (9)) will reduce the computational cost. We performed some experiments with this loss function. Although with this loss function the training time is almost equal to that of f_{DSC}, the segmentation results were not very encouraging. For example on the Pancreas CT dataset, HD was 24.8 ± 11.9, which was statistically significantly larger than that of Loss_{DT}. Nonetheless, in our experiments, this low-cost loss function always reduced the HD compared with f_{DSC}.

Although the results with $f_{\text{HD}}^{\text{ER}}$ were slightly worse than with $f_{\text{HD}}^{\text{DT}}$ and $f_{\text{HD}}^{\text{CV}}$, it still significantly reduced HD compared with f_{DSC}. Moreover, it adds little computational overhead to a CNN. The likely reason for the lower performance of $f_{\text{HD}}^{\text{ER}}$ is that it is not as accurate as the other two methods in estimating HD. On some images, it can greatly underestimate HD. Our approach to estimating HD using morphological operations is a simple one. In addition to the cross-shaped structuring elements described above, we also experimented with square-shaped and cube-shaped elements (for 2D and 3D, respectively). However, the results in terms of the spread of the points in Figure 6 and also in terms of segmentation accuracy were not better than those obtained with the cross-shaped elements. Nonetheless, it may be possible to design more accurate, but still fast and simple, methods based on morphological operations proposed in [39].

For Loss_{CV}, the choice of the set of radius values, R, can be considered as a hyper-parameter, which should be chosen for each application based on the expected range of segmentation errors. As we have mentioned above, the choice of R will affect both the segmentation accuracy and computational time. To give the reader a sense of this trade-off, in Table II we have shown the results of some experiments on 2D prostate ultrasound and 3D Pancreas CT with different R. Some conclusions can be drawn from these results. First, as we speculated in Section II-D, no gain is achieved by using kernels of every size. This can be seen by comparing the results obtained with $R = \{3, 6, 9, \ldots, 18\}$ versus $R = \{3, 4, 5, \ldots, 18\}$ for 2D prostate ultrasound dataset and by comparing the results obtained with $R = \{3, 6, \ldots, 15\}$ versus $R = \{3, 5, \ldots, 15\}$ for 3D Pancreas CT dataset. Using more tightly-packed values of R will only increase the computational cost without substantially improving the segmentation results. The results in Table II also show that the segmentation performance is not very sensitive to the choice of R. In particular, for both datasets used in these experiments, all choices of R resulted in much smaller HD than with $f_{\text{DSC}}(q,p)$ (as shown in Table I). Table I above also showed that the same choice of $R = \{3, 6, 9\}$ led to very good results on three different datasets of 3D prostate MRI, 3D liver CT and 3D pancreas CT. This is further evidence that this method is not highly sensitive to the choice

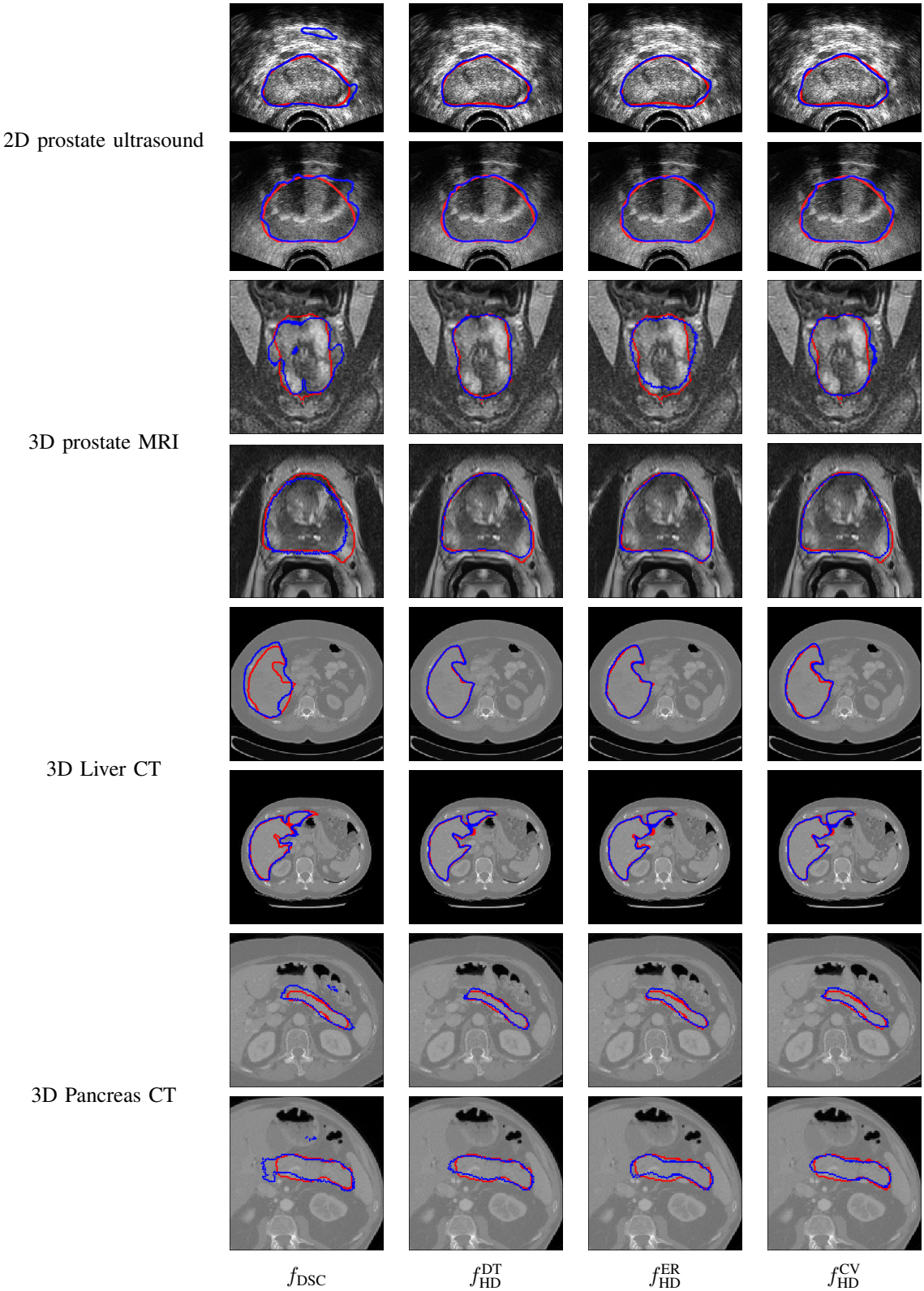


Figure 9. Selected images from each dataset and the boundaries of the segmentations produced by different loss functions (in blue) and the ground-truth segmentation (in red). For the 3D images, we have shown the slice on which f_{DSC} had the largest segmentation error.

of R.

Our proposed DT-based loss function, $f_{\text{HD}}^{\text{DT}}$, is based on a weighting of the segmentation errors, where larger distance errors are weighted more strongly. This approach seems to be the opposite of what some previous studies have proposed. For example, for cell segmentation in [29] and for prostate segmentation in [30], it has been suggested that larger weights be assigned to the pixels that are closer to the boundary of the

Table II
THE RESULTS OF EXPERIMENTS WITH VARIOUS CHOICES OF THE PARAMETER R IN Loss_{CV} ON TWO DIFFERENT DATASETS. FOR EACH DATASET, THE FIRST ROW SHOWS THE RESULT WITH OUR DEFAULT CHOICE OF R WHICH WAS PRESENTED IN TABLE I. THE LAST ROW FOR EACH DATASET SHOWS THE RESULT OBTAINED WITH f_{DSC} (ALSO FROM TABLE I) FOR EASY COMPARISON.

Dataset	R	DSC	HD (mm)	Training time (h)
2D prostate ultrasound	$R = \{3, 6, 9, \ldots, 18\}$	0.941 ± 0.036	2.7 ± 1.8	4.1
	$R = \{3, 6, 9, \ldots, 30\}$	0.943 ± 0.030	2.7 ± 1.6	9.8
	$R = \{3, 6, 9\}$	0.936 ± 0.045	3.0 ± 2.2	3.8
	$R = \{3, 4, 5, \ldots, 18\}$	0.940 ± 0.035	2.7 ± 1.8	12.0
	$R = \{3, 4, 5, \ldots, 30\}$	0.944 ± 0.032	2.6 ± 1.6	20.4
	$f_{\text{DSC}}(q, p)$	0.932 ± 0.039	4.3 ± 2.8	3.2
3D Pancreas CT	$R = \{3, 6, 9\}$	0.780 ± 0.055	21.7 ± 11.0	34
	$R = \{3, 6, \ldots, 15\}$	0.790 ± 0.056	21.2 ± 10.7	51
	$R = \{3, 6\}$	0.772 ± 0.082	24.0 ± 12.8	28
	$R = \{3, 5, \ldots, 9\}$	0.779 ± 0.050	21.4 ± 10.6	40
	$R = \{3, 5, \ldots, 15\}$	0.788 ± 0.061	21.3 ± 10.1	63
	$f_{\text{DSC}}(q, p)$	0.752 ± 0.120	32.1 ± 17.0	22

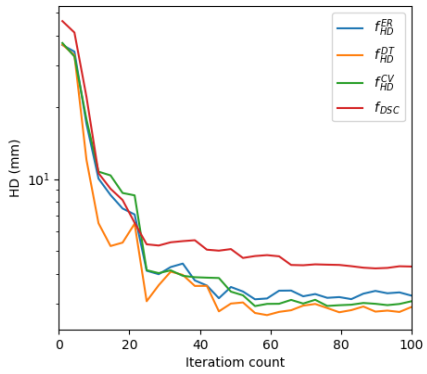


Figure 10. A plot of the mean HD on the test data for the 2D TRUS images of prostate as a function of the training epoch number for different loss functions.

ground-truth segmentation. To test this alternate approach, we used the following loss function, which is similar to the loss function suggested in [29]:

$$\text{Loss}_{\text{DT}}^{\dagger}(q, p) = \frac{1}{|\,\Omega\,|} \sum_{\Omega} \left((p - q)^2 \circ \exp\left(\frac{-d_p^2}{2\sigma^2} \right) \right) \quad (18)$$

Our observations show that although this loss function may slightly improve the DSC on some datasets, in general it has no significant positive effect on HD, which is the focus of this work. For example, with the above loss function (with an added DSC loss term as in Equation (17)), on the 2D TRUS prostate data we achieved DSC and HD of 0.938 ± 0.035 and 4.0 ± 2.7 mm, respectively. Paired t-tests showed that HD was significantly larger than our three HD-based loss functions and DSC was not significantly different from those obtained with other loss functions on this dataset. Therefore, compared with our proposed loss functions, assigning larger weights to the pixels closer to the ground-truth boundary harms the segmentation performance in terms of HD. It is worth pointing out that the main challenge in [29] was to segment the boundaries of the object of interest (cells), which could justify a loss function as in Equation (18). Moreover, unlike [29], [30], our loss functions include a DSC loss term, which means that the comparison of our results with those studies is not quite fair.

Another distinct aspect of our proposed HD-based loss functions is that they are based on the squared difference of the probability maps, i.e., $(p - q)^2$, whereas medical image segmentation methods typically work with the cross-entropy. Our choice of the ℓ_2-norm was motivated by the results reported in some recent studies [33], [47], [48]. These studies have shown that loss functions based on hinge loss, squared hinge loss, and ℓ_1 and ℓ_2 norms may lead to superior results in different deep learning models. Inspired by these studies, and because none of them had considered the application of image segmentation, we conducted a set of experiments to examine the usefulness of these formulations in our application. Figure 11 shows an example of our observations. In this experiment, we replaced the squared difference term $(p - q)^2$ in the DT-based loss function (Equation (8)) with some of the alternatives proposed in [33], [47], [48]. As can be seen in this figure, ℓ_2 loss, hinge loss, and squared hinge loss all perform better than the cross-entropy loss, which is widely used in CNN-based image segmentation methods. It was based on such observations that we built our HD-based loss functions (Equations (8), (12), and (14)) upon the squared difference, $(p - q)^2$. Overall, our observation are in line with those reported in [33]. However, we observed that the ℓ_2 loss gives slightly better results than the hinge and squared hinge losses and that the ℓ_1 loss is not very poor either, whereas the authors of [33] found that the ℓ_1 loss was very poor and the hinge losses were slightly better than the ℓ_2 loss. We think that the most likely reason for these differences is the extra DSC loss term in our work. Moreover, our image segmentation problem is quite different from the applications considered in [33].

Overall, the results reported in Table I are close to or better than the results reported by many recent studies. On the 2D TRUS prostate image data, our results in terms of HD are much better than those reported in [49], [50] on the same dataset, where the authors have reported HD of above 5 mm

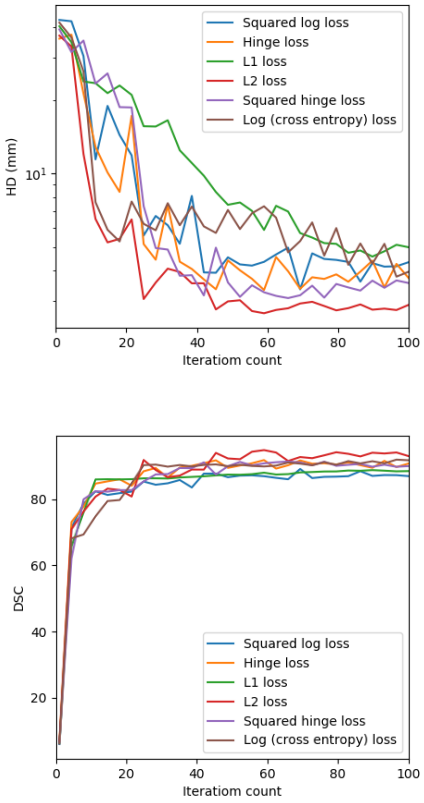


Figure 11. Plots of Dice Similarity Coefficient and Hausdorff Distance on the 2D TRUS images of the prostate for different formulations of the distance transform-based loss function.

using two different methods. For prostate segmentation in 3D MRI, most studies have only reported DSC. Some studies have reported the 95th-percentile of HD within an image [31], [51], [52]. Note that this is different from the *inter-patient* 90th-percentile of HD that we have reported in this paper. The mean of the 95th-percentile intra-patient HD reported in a recent comparison of several state of the art methods is in the range $4.9 - 7.6$ mm [52], whereas this quantity computed on the test data with the model trained using $f_{\text{HD}}^{\text{DT}}(q,p)$ in our work is 4.70 ± 0.97 mm. Incidentally, the lowest HD in the comparison published in [52] was achieved by a method from our own group that has not been published yet. For liver segmentation in CT, the reported values of HD vary greatly between 24 mm and 119 mm [53]. Our best result, as can be seen in Table I is 25.1 mm. For pancreas segmentation, most studies only report DSC and the mean or root-mean-square of surface distance. A recent study reported values of HD in the range $17.7 - 22.2$ mm [54], compared with our best result of 21.3 mm. Pancreas segmentation is considered to be very challenging and most studies have reported a DSC of below 0.80. Recently some works have achieved DSC values of well above 0.80 [55], [56]. We should note, however, that those studies have used more elaborate machinery such as a separate model to identify the location of the pancreas or iterative refinement strategies. For example, the work that has reported mean HDs as low as 17.7 mm is a multi-stage method that uses multiple CNN models and other machine learning methods to localize the pancreas, identify boundary cues, and aggregate segmentation cues [54]. We should stress that the goal of the present study was not to achieve the state of the art results in segmentation of different organs and imaging modalities; one could always achieve better results by fine-tuning the network structure or employing a more sophisticated methodology that is tailored to the specific organ and imaging modality. Our experiments were intended to show that the proposed loss functions could lead to a significant reduction in HD. That is why we have adopted standard segmentation models and training procedures in order to eliminate, as much as possible, other confounding factors and study the effectiveness of the proposed loss functions.

We further compared our proposed loss functions on the test data of the PROMISE12 challenge [43]. In this experiment, we trained our CNN model with different loss functions on the 50 training images provided by this challenge and tested on the 30 test images for which the true segmentation is only known to the challenge organizers. A summary of the results of this experiment is presented in Table III. The trends observed in this Table are similar to those in Table I. The DSC scores achieved on this dataset are better than those in Table I and they are very similar for all four loss functions. Compared with f_{DSC}, all three proposed HD-based loss functions have substantially reduced the mean, standard deviation, and the maximum of HD95. The reduction in HD for 3D prostate MRI data in Table I was in the range $18 - 23\%$, whereas the reduction in HD95 in III was in the range $10 - 15\%$. This is, at least in part, because HD95 ignores the top 5% with the largest surface distance error. Therefore HD95 does not reflect the very largest segmentation errors that HD represent. We cannot compute the HD for this dataset because we do not have the ground-truth. Nonetheless, these results show that our proposed method is capable of reducing not only the very largest segmentation error but also to consistently reduce large segmentation errors as quantified by HD95. Compared with other recently-published papers on the same dataset, our achieved HD95 values are significantly better. Two recently-published methods evaluated on the same dataset have been included in Table III. As we pointed out above, a recent study compared 10 state of the art methods on this dataset and reported the HD95 values in the range $4.9 - 7.6$ mm [52], whereas our HD-based loss functions resulted in HD values as low as 4.26.

We also applied our methods on the MRI brain segmentation data from the iSeg-2017 challenge [58]. The goal of this challenge is to segment infant brain MR images into four classes: 1) white matter (WM), 2) gray matter (GM), 3) cerebrospinal fluid (CSF), and 4) background. The challenge organizers allow a maximum of two submissions per team. Because based on Table I $f_{\text{HD}}^{\text{DT}}$ and $f_{\text{HD}}^{\text{ER}}$ were, overall, the best and worst of our three HD-based loss functions, we evaluated these two loss functions on this challenge. Our results have been shown in Table IV. As a comparison with another method, we have included the results obtained by the recently published method in [59], which at the time of our participation in this challenge (March 2019) has achieved the highest Dice score on all three

Table III
A SUMMARY OF THE COMPARISON OF THE PROPOSED LOSS FUNCTIONS ON THE TEST DATA FROM THE PROMISE12 CHALLENGE [43]. N.R. STANDS FOR "NOT REPORTED".

Loss Function	DSC	HD95 (mm)	Max. of HD95 (mm)
$f_{\text{DSC}}(q,p)$	$\mathbf{0.908 \pm 0.032}$	5.00 ± 2.16	11.3
$f_{\text{HD}}^{\text{DT}}(q,p)$	0.902 ± 0.026	4.28 ± 1.05	7.8
$f_{\text{HD}}^{\text{ER}}(q,p)$	0.904 ± 0.023	4.48 ± 1.46	8.8
$f_{\text{HD}}^{\text{CV}}(q,p)$	0.902 ± 0.025	$\mathbf{4.26 \pm 1.03}$	**7.6**
Yu et al., 2017, [57]	0.894	5.54	N.R.
Brosch et al., 2018, [52]	0.905	4.94	N.R.

structures among all participating teams. To also compare our other two loss functions (f_{DSC} and $f_{\text{HD}}^{\text{CV}}$) we performed a five-fold cross-validation on the training images of this challenge. The results of this experiment have also been included in Table IV.

As can be seen from this table, $f_{\text{HD}}^{\text{DT}}$ and $f_{\text{HD}}^{\text{ER}}$ achieve a lower HD than the method of [59] on all three structures, with one exception of loss function $f_{\text{HD}}^{\text{ER}}(q,p)$ on CSF segmentation). Our best performing loss function ($f_{\text{HD}}^{\text{DT}}(q,p)$) has reduced HD compared with [59] by approximately 1.5%, 35%, and 4.2%, respectively on CSF, GM, and WM, respectively. Overall, compared with all participating teams in this challenge, our best achieved HD was ranked 1st in CSF segmentation, 3rd in GM segmentation, and 6th in WM segmentation. We should note that we used the basic 3D U-Net that we had used for segmentation of the other datasets in this study, without any modifications. On the other hand, other participating teams have developed methods specifically for this challenge. For example, the leading method in [59] has used a CNN-based patch-level training and prediction aggregation method specially designed for this brain MRI segmentation task. A better way of assessing the effect of our proposed HD-based loss functions is to compare with the results obtained with f_{DSC} in our five-fold cross-validation experiments reported in the same Table. Compared with f_{DSC}, our best performing loss function, $f_{\text{HD}}^{\text{DT}}$, reduces HD by 15%, 35%, and 23%, respectively on CSF, GM, and WM, respectively, which represent substantial improvements. The other two HD-based loss functions, $f_{\text{HD}}^{\text{ER}}$ and $f_{\text{HD}}^{\text{CV}}$, also reduce the HD (compared with f_{DSC}) by 8% to 33% while achieving DSC values that are very close to or better than f_{DSC}.

As we have argued throughout the paper, minimizing HD directly can be tricky and counter-productive. This is why we have proposed relaxed loss functions that are based on HD, instead of minimizing HD directly. Moreover, we augmented our HD-based loss functions with DSC loss, which is based on the amount of overlap between the ground-truth and estimated segmentation maps. Nonetheless, one would be curious to know how the proposed HD-based loss functions would perform without the added DSC loss term. We performed experiments to empirically understand the training of CNN segmentation methods with these loss functions. We observed that it was possible to train CNN segmentation methods with these loss functions. However, this required a more careful tuning of the learning rate and using a much smaller learning rate in the first few training epochs. Moreover, we observed that the segmentation results were always better when the DSC loss term was included.

Here we briefly summarize the results obtained on two of our datasets. For the 2D prostate ultrasound data, the HD values achieved using $f_{\text{HD}}^{\text{DT}}$, $f_{\text{HD}}^{\text{ER}}$, and $f_{\text{HD}}^{\text{CV}}$ without the DSC loss term were, respectively, 2.9 ± 1.8, 3.3 ± 2.4, and 2.9 ± 2.2 and the DSC values were, respectively, 0.930 ± 0.038, 0.921 ± 0.048, and 0.923 ± 0.040. These results are slightly worse than those reported in Table I for these methods with the added DSC loss term. Nonetheless, it is interesting to note that these HD values are still much better than 4.3 ± 2.8 achieved with f_{DSC} as shown in Table I. For the 3D pancreas CT data, the HD values achieved using $f_{\text{HD}}^{\text{DT}}$, $f_{\text{HD}}^{\text{ER}}$, and $f_{\text{HD}}^{\text{CV}}$ without the DSC loss term were, respectively, 22.9 ± 13.8, 27.6 ± 13.4, and 22.7 ± 12.0 and the DSC values were, respectively, 0.779 ± 0.058, 0.750 ± 0.068, and 0.772 ± 0.054. These results also show that better results can be obtained by including the DSC loss term. Nonetheless, these HD values are still much smaller than 32.1 ± 17.0 obtained with f_{DSC} as shown in Table I.

Using the HD-based loss functions alone is equivalent to setting $\lambda = 0$ in Eq. 17. Using the DSC loss alone, i.e., f_{DSC} corresponds to a very large λ. As we mentioned above, in our experiments we update the value of λ such that the two loss terms (i.e., the HD-based loss term and the DSC loss term) have equal weight. In Figure 12 we have shown the effect of changing this weighting on the example of training with $f_{\text{HD}}^{\text{DT}}$ on the 2D prostate ultrasound data. This figure shows that for this specific case this choice is close to optimal and that choosing λ such that the ratio of the DSC loss term to the HD-based loss term is in the range $[0.1, 2]$ leads to good segmentation results in terms of both HD and DSC. Figure 13 shows the change in the value of λ over training epochs for the three HD-based loss functions on the 3D liver CT data. We observed similar trends on the other datasets. The values of λ have been normalized such that the value at the end of training is approximately equal to one, so that the curves can be shown on the same plot. The overall trend is that the value of λ in the early training epochs is larger. This is because at the start of training there are many false positives far from the segmentation boundaries, which increases the HD-based loss term.

We further tried training with the "exact HD" as the loss function, instead of our proposed relaxed HD-based loss functions. In these experiments, we tried training models using Equation (7) as the loss function for 2D and 3D datasets.

Table IV
A SUMMARY OF THE COMPARISON OF THE PROPOSED LOSS FUNCTIONS ON THE TEST DATA FROM THE ISEG-2017 CHALLENGE.

		CSF		GM		WM	
Experiment	Loss Function	DSC	HD	DSC	HD	DSC	HD
Evaluated on the test data of iSeg2017	$f_{\text{HD}}^{\text{DT}}(q,p)$	0.945	**8.720**	0.911	**6.212**	0.890	**6.812**
	$f_{\text{HD}}^{\text{ER}}(q,p)$	0.947	9.434	0.915	6.832	0.894	6.988
	Hashemi et al., 2019 [59]	**0.960**	8.850	**0.926**	9.557	**0.907**	7.104
Evaluated on the training data of iSeg2017 via five-fold cross-validation	$f_{\text{DSC}}(q,p)$	0.950	10.224	0.914	9.320	0.905	8.802
	$f_{\text{HD}}^{\text{DT}}(q,p)$	0.942	**8.725**	0.910	**6.101**	0.894	**6.816**
	$f_{\text{HD}}^{\text{ER}}(q,p)$	0.949	9.408	0.910	6.845	0.899	6.983
	$f_{\text{HD}}^{\text{CV}}(q,p)$	**0.951**	8.733	**0.922**	6.225	**0.908**	6.854

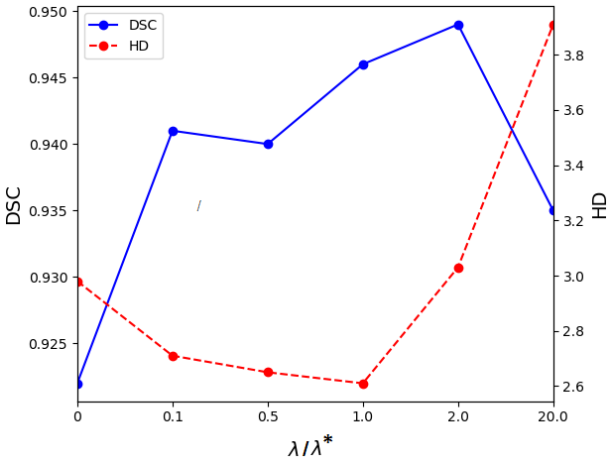


Figure 12. The change in the segmentation performance on the 2D prostate ultrasound data using loss function $f_{\text{HD}}^{\text{DT}}$ with different values of λ. The horizontal axis is a function of λ/λ^*, where λ^* is our default setting which is chosen such that the HD-based and DSC loss terms have equal weights.

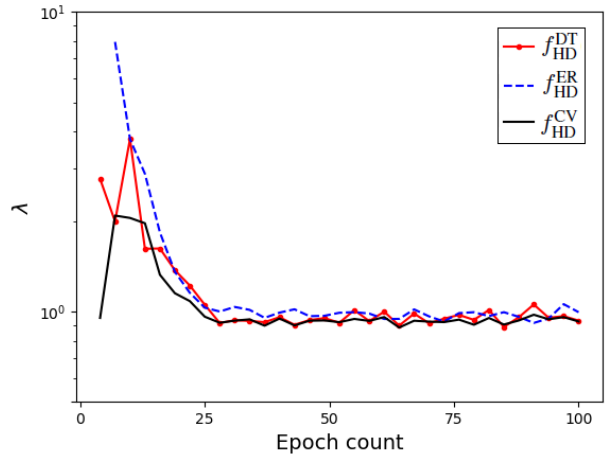


Figure 13. Change in the value of λ with training epochs for the 3D liver CT dataset. The values of λ for the three loss functions have been normalized such that the value at the end of training is approximately equal to one.

However, the models never converged to a meaningful state regardless of the learning rate. This is very much expected because HD aims at minimizing the error at one single point with the largest error. We have presented arguments against using HD as a loss function in Section I and reviewed some of the computer vision literature in this regard. Moreover, with randomly-initialized deep learning models, minimizing only the largest error cannot be justified and our observations confirm this.

We have successfully tested our method on five different datasets which include organs of different sizes and shapes in different imaging modalities. Nonetheless, we cannot claim that our method will be successful in all medical image segmentation tasks. Indeed, the range of anatomical shapes and the level of detail in medical image segmentation is very wide. Hence, application of our loss functions to other organs may need modifications. For example, in some applications such as vessel segmentation, recent studies have used additional modules such as probabilistic graphical models on top of CNNs to achieve good results [60], [61]. In such applications, our proposed methods might need application-specific modifications. A comprehensive experimental study or review of the range of applications for our proposed method and a discussion of all application-specific issues is beyond the scope of this paper.

IV. CONCLUSION

Our results show that all three proposed HD-based loss functions can lead to statistically significant reductions in HD in the segmentation of 2D and 3D medical images of different imaging modalities. Therefore, the proposed methods can be very useful in applications such as multimodal image registration where large segmentation errors can be very harmful. To the best of our knowledge, none of the three methods proposed in this work for estimating HD and the three loss functions proposed for training segmentation algorithms have appeared in previous publications. The distance transform-based loss, $f_{\text{HD}}^{\text{DT}}$, is the most intuitive of the three formulations and leads to very good results, but it also substantially increases the computational load. The loss based on morphological erosion, $f_{\text{HD}}^{\text{ER}}$, is computationally less expensive, but not as effective as the other two losses. The loss based on convolutions with kernels of increasing sizes, $f_{\text{HD}}^{\text{CV}}$, leads to very good results and it can be computationally much less demanding than the distance transform-based loss if one properly limits the maximum size and the number of convolutional kernels that are used. Overall, the decision on which of the three loss functions to use depends on the application. For example, for 2D images the cost of computing the distance transforms is not substantial and one may use $f_{\text{HD}}^{\text{DT}}$. For 3D images, $f_{\text{HD}}^{\text{ER}}$ and $f_{\text{HD}}^{\text{CV}}$

may be better options, with $f_{\mathrm{HD}}^{\mathrm{CV}}$ offering better segmentation performance at the cost of longer training times.

To the best of our knowledge, this is the first work to aim at reducing HD in medical image segmentation. The methods presented in this paper may be improved in several ways. Faster implementation of the HD-based loss functions and more accurate implementation of the loss function based on morphological erosion would be useful. Moreover, extension of the methods for other applications such as vessel segmentation could also be pursued.

ACKNOWLEDGMENT

This project is funded by the Natural Science and Engineering Research Council of Canada (NSERC), the Canadian Institutes of Health Research (CIHR), and the Prostate Cancer Canada (PCC). We would like to thank the support from the Charles Laszlo Chair in Biomedical Engineering held by Professor S. Salcudean.

REFERENCES

[1] K. D. Toennies, *Guide to Medical Image Analysis*. Springer, 2017.

[2] S. K. Zhou, *Medical Image Recognition, Segmentation and Parsing: Machine Learning and Multiple Object Approaches*. Academic Press, 2015.

[3] H. R. Roth, C. Shen, H. Oda, M. Oda, Y. Hayashi, K. Misawa, and K. Mori, "Deep learning and its application to medical image segmentation," *Medical Imaging Technology*, vol. 36, no. 2, pp. 63–71, 2018.

[4] G. Wang, W. Li, M. A. Zuluaga, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, and T. Vercauteren, "Interactive medical image segmentation using deep learning with image-specific fine tuning," *IEEE Transactions on Medical Imaging*, vol. 37, no. 7, pp. 1562–1573, July 2018.

[5] W. R. Crum, O. Camara, and D. L. Hill, "Generalized overlap measures for evaluation and validation in medical image analysis," *IEEE Transactions on Medical Imaging*, vol. 25, no. 11, pp. 1451–1461, 2006.

[6] A. A. Taha and A. Hanbury, "Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, p. 29, 2015.

[7] R. T. Rockafellar and R. J.-B. Wets, *Variational Analysis*. Springer Science & Business Media, 2009, vol. 317.

[8] E. Smistad, T. L. Falch, M. Bozorgi, A. C. Elster, and F. Lindseth, "Medical image segmentation on GPUs–a comprehensive review," *Medical image analysis*, vol. 20, no. 1, pp. 1–18, 2015.

[9] T. Heimann and H.-P. Meinzer, "Statistical shape models for 3d medical image segmentation: A review," *Medical Image Analysis*, vol. 13, no. 4, pp. 543 – 563, 2009.

[10] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, 1997.

[11] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, 1988.

[12] J. L. Marroquin, B. C. Vemuri, S. Botello, E. Calderon, and A. Fernandez-Bouzas, "An accurate and efficient bayesian method for automatic segmentation of brain MRI," *IEEE Transactions on Medical Imaging*, vol. 21, no. 8, pp. 934–945, 2002.

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

[14] N. Makni, N. Betrouni, and O. Colot, "Introducing spatial neighbourhood in evidential c-means for segmentation of multi-source images: Application to prostate multi-parametric mri," *Information Fusion*, vol. 19, pp. 61–72, 2014.

[15] S. Pereira, A. Pinto, J. Oliveira, A. M. Mendrik, J. H. Correia, and C. A. Silva, "Automatic brain tissue segmentation in mr images using random forests and conditional random fields," *Journal of Neuroscience Methods*, vol. 270, pp. 111–123, 2016.

[16] F. R. Schmidt and Y. Boykov, "Hausdorff distance constraint for multi-surface segmentation," in *European Conference on Computer Vision*. Springer, 2012, pp. 598–611.

[17] M.-P. Dubuisson and A. K. Jain, "A modified hausdorff distance for object matching," in *Proceedings of 12th International Conference on Pattern Recognition*. IEEE, pp. 566–568.

[18] C.-H. T. Yang, S.-H. Lai, and L.-W. Chang, "Hybrid image matching combining hausdorff distance with normalized gradient matching," *Pattern Recognition*, vol. 40, no. 4, pp. 1173–1181, 2007.

[19] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[20] K. J. Kirchberg, O. Jesorsky, and R. W. Frischholz, "Genetic model optimization for hausdorff distance-based face localization," in *International Workshop on Biometric Authentication*. Springer, 2002, pp. 103–111.

[21] D.-G. Sim, O.-K. Kwon, and R.-H. Park, "Object matching algorithms using robust hausdorff distance measures," *IEEE Transactions on Image Processing*, vol. 8, no. 3, pp. 425–429, 1999.

[22] H. Tan and Y.-J. Zhang, "A novel weighted hausdorff distance for face localization," *Image and Vision Computing*, vol. 24, no. 7, pp. 656–662, 2006.

[23] C. Knauer, K. Kriegel, and F. Stehn, "Minimizing the weighted directed hausdorff distance between colored point sets under translations and rigid motions," *Theoretical Computer Science*, vol. 412, no. 4-5, pp. 375–382, 2011.

[24] D. Dimitrov, C. Knauer, K. Kriegel, and F. Stehn, "Approximation algorithms for a point-to-surface registration problem in medical navigation," in *International Workshop on Frontiers in Algorithmics*. Springer, 2007, pp. 26–37.

[25] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, no. Supplement C, pp. 60 – 88, 2017.

[26] F. Milletari, A. Rothberg, J. Jia, and M. Sofka, *Integrating Statistical Prior Knowledge into Convolutional Neural Networks*. Cham: Springer International Publishing, 2017, pp. 161–168.

[27] D. Karimi, G. Samei, C. Kesch, G. Nir, and S. E. Salcudean, "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models," *International Journal of Computer Assisted Radiology and Surgery*, vol. 13, no. 8, pp. 1211–1219, Aug 2018.

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[29] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2015, pp. 234–241.

[30] E. M. A. Anas, S. Nouranian, S. S. Mahdavi, I. Spadinger, W. J. Morris, S. E. Salcudean, P. Mousavi, and P. Abolmaesumi, "Clinical target-volume delineation in prostate brachytherapy using residual neural networks," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 365–373.

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

[32] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3d fully convolutional deep networks," in *International Workshop on Machine Learning in Medical Imaging*. Springer, 2017, pp. 379–387.

[33] K. Janocha and W. M. Czarnecki, "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.

[34] R. Szeliski, *Computer Vision: Algorithms and Applications*. Springer Science & Business Media, 2010.

[35] T. Jech, "Set theory: The third millennium edition, revised and expanded, 3rd," *Springer Monographs in Mathematics, Springer-Verlag Berlin Heidelberg New York*, 2002.

[36] A. A. Taha and A. Hanbury, "An efficient algorithm for calculating the exact hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 11, pp. 2153–2163, 2015.

[37] P. Felzenszwalb and D. Huttenlocher, "Distance transforms of sampled functions," Cornell University, Tech. Rep., 2004.

[38] C. R. Maurer, R. Qi, and V. Raghavan, "A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 265–270, 2003.

[39] J. Mazille, "Mathematical morphology and convolutions," *Journal of Microscopy*, vol. 156, no. 1, pp. 3–13, 1989.

[40] M. Razaz and D. Hagyard, "Efficient convolution based algorithms for erosion and dilation." in *NSIP*, 1999, pp. 360–363.

[41] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, 2013.

[42] S. S. Mahdavi, N. Chng, I. Spadinger, W. J. Morris, and S. E. Salcudean, "Semi-automatic segmentation for prostate interventions," *Medical Image Analysis*, vol. 15, no. 2, pp. 226–237, 2011.

[43] G. Litjens, R. Toth, W. van de Ven, C. Hoeks, S. Kerkstra, B. van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang, R. Strand, F. Malmberg, Y. Ou, C. Davatzikos, M. Kirschner, F. Jung, J. Yuan, W. Qiu, Q. Gao, P. Edwards, B. Maan, F. van der Heijden, S. Ghose, J. Mitra, J. Dowling, D. Barratt, H. Huisman, and A. Madabhushi, "Evaluation of prostate segmentation algorithms for mri: The promise12 challenge," *Medical Image Analysis*, vol. 18, no. 2, pp. 359 – 373, 2014.

[44] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, "N4itk: Improved n3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, June 2010.

[45] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016, pp. 424–432.

[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*, 2014.

[47] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[48] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.

[49] S. Nouranian, S. S. Mahdavi, I. Spadinger, W. J. Morris, S. E. Salcudean, and P. Abolmaesumi, "A multi-atlas-based segmentation framework for prostate brachytherapy," *IEEE Transactions on Medical Imaging*, vol. 34, no. 4, pp. 950–961, 2015.

[50] S. Nouranian, M. Ramezani, I. Spadinger, W. J. Morris, S. E. Salcudean, and P. Abolmaesumi, "Learning-based multi-label segmentation of transrectal ultrasound images for prostate brachytherapy," *IEEE Transactions on Medical Imaging*, vol. 35, no. 3, pp. 921–932, 2016.

[51] A. Salimi, M. A. Pourmina, and M.-S. Moin, "Fully automatic prostate segmentation in mr images using a new hybrid active contour-based approach," *Signal, Image and Video Processing*, pp. 1–9, 2018.

[52] T. Brosch, J. Peters, A. Groth, T. Stehle, and J. Weese, "Deep learning-based boundary detection for model-based segmentation with application to mr prostate segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, A. F. Frangi, J. A. Schnabel, C. Davatzikos, C. Alberola-López, and G. Fichtinger, Eds. Cham: Springer International Publishing, 2018, pp. 515–522.

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

[54] H. R. Roth, L. Lu, N. Lay, A. P. Harrison, A. Farag, A. Sohn, and R. M. Summers, "Spatial aggregation of holistically-nested convolutional neural networks for automated pancreas localization and segmentation," *Medical Image Analysis*, vol. 45, pp. 94–107, 2018.

[55] H. Roth, M. Oda, N. Shimizu, H. Oda, Y. Hayashi, T. Kitasaka, M. Fujiwara, K. Misawa, and K. Mori, "Towards dense volumetric pancreas segmentation in ct using 3d fully convolutional networks," in *Medical Imaging 2018: Image Processing*, vol. 10574. International Society for Optics and Photonics, 2018, p. 105740B.

[56] Y. Zhou, L. Xie, W. Shen, Y. Wang, E. K. Fishman, and A. L. Yuille, "A fixed-point model for pancreas segmentation in abdominal ct scans," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 693–701.

[57] L. Yu, X. Yang, H. Chen, J. Qin, and P. A. Heng, "Volumetric convnets with mixed residual connections for automated prostate segmentation from 3d MR images," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[58] L. Wang, D. Nie, G. Li, É. Puybareau, J. Dolz, Q. Zhang, F. Wang, J. Xia, Z. Wu, J. Chen *et al.*, "Benchmark on automatic 6-month-old infant brain segmentation algorithms: The iseg-2017 challenge," *IEEE Transactions on Medical Imaging*, 2019.

[59] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, and A. Gholipour, "Asymmetric loss functions and deep densely-connected networks for highly-imbalanced medical image segmentation: Application to multiple sclerosis lesion detection," *IEEE Access*, vol. 7, pp. 1721–1735, 2019.

[60] H. Fu, Y. Xu, D. W. K. Wong, and J. Liu, "Retinal vessel segmentation via deep learning network and fully-connected conditional random fields," in *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*. IEEE, 2016, pp. 698–701.

[61] S. Y. Shin, S. Lee, I. D. Yun, and K. M. Lee, "Deep vessel segmentation by learning graphical connectivity," *arXiv preprint arXiv:1806.02279*, 2018.