Combo Loss: Handling Input and Output Imbalance in Multi-Organ Segmentation

Saeid Asgari Taghanaki*[a,b], Yefeng Zheng[b], S. Kevin Zhou[b], Bogdan Georgescu[b], Puneet Sharma[b], Daguang Xu[b], Dorin Comaniciu[b], Ghassan Hamarneh[a]

[a]*School of Computing Science, Simon Fraser University, Canada*
[b]*Medical Imaging Technologies, Siemens Healthineers, Princeton, NJ, USA*

Abstract

Simultaneous segmentation of multiple organs from different medical imaging modalities is a crucial task as it can be utilized for computer-aided diagnosis, computer-assisted surgery, and therapy planning. Thanks to the recent advances in deep learning, several deep neural networks for medical image segmentation have been introduced successfully for this purpose. In this paper, we focus on learning a deep multi-organ segmentation network that labels voxels. In particular, we examine the critical choice of a loss function in order to handle the notorious imbalance problem that plagues both the input and output of a learning model. The input imbalance refers to the class-imbalance in the input training samples (i.e., small foreground objects embedded in an abundance of background voxels, as well as organs of varying sizes). The output imbalance refers to the imbalance between the false positives and false negatives of the inference model. In order to tackle both types of imbalance during training and inference, we introduce a new curriculum learning based loss function. Specifically, we leverage Dice similarity coefficient to deter model parameters from being held at bad local minima and at the same time gradually learn better model parameters by penalizing for false positives/negatives using a cross entropy term. We evaluated the proposed loss function on three datasets: whole body positron emission tomography (PET) scans with 5 target organs, magnetic resonance imaging (MRI) prostate scans, and ultrasound echocardigraphy images with a single target organ i.e., left ventricular. We show that a simple network architecture with the proposed integrative loss function can outperform state-of-the-art methods and results of

*Corresponding author

Email addresses: sasagarit@sfu.ca (Saeid Asgari Taghanaki*), yefeng.zheng@gmail.com (Yefeng Zheng), s.kevin.zhou@gmail.com (S. Kevin Zhou), bogdan.georgescu@siemens-healthineers.com (Bogdan Georgescu), sharma.puneet@siemens-healthineers.com (Puneet Sharma), daguang.xu@siemens-healthineers.com (Daguang Xu), dorin.comaniciu@siemens.com (Dorin Comaniciu), hamarneh@sfu.ca (Ghassan Hamarneh)

the competing methods can be improved when our proposed loss is used.

Keywords: Class-imbalance, output imbalance, deep convolutional neural networks, loss function, multi-organ segmentation

1. Introduction

Organ segmentation is an important processing step in medical image analysis, e.g., for image guided interventions, radiotherapy, or improved radiological diagnostics. A plethora of single/multi-organ segmentation methods including machine/deep learning approaches has been introduced in the literature for different medical imaging modalities, e.g., magnetic resonance imaging, and positron emission tomography (PET).

More recently, deep learning based medical image segmentation approaches have gained great popularity [1–6]. Several deep convolutional segmentation models in the form of encoder-decoder networks have been proposed for both medical and non-medical images to learn features and classify/segment images simultaneously in an end-to-end manner e.g., 2D U-Net [7], 3D U-Net [8], 3D V-Net [9], 2D SegNet [10]. These models with/without modifications have been widely applied to both binary and multi-class medical image segmentation problems.

When performing segmentation especially using deep networks, one has to cope with two types of imbalance issues:

a) Input imbalance (i.e., prior probability) or inter-class-imbalance during training, i.e., much fewer foreground pixels/voxels relative to the large number of background voxels in binary segmentation, and smaller objects/classes in a multi-class segmentation relative to other larger objects/classes and the background. Therefore, classes with more observations (i.e., voxels) overshadow the minority classes.

b) Output imbalance (i.e., posterior probability). During inference, it is unavoidable to have false positives and false negatives. False positives are the background voxels (or other objects in the case of multi-class) that are wrongly labeled as the target object. False negatives refer to the voxels of a target object that are erroneously labeled as background or, in the case of multi-organ segmentation, mislabelled as another organ. Clearly, eliminating both false positives and false negatives is the ultimate ideal. However, in practical systems, one increases as the other decreases. For certain applications, reducing the false positive (FP) rate is more important than reducing the false negative

(FN) rate or vice versa. The following are example cases where FP should be penalized more: to handle missing organs and to prevent a model from segmenting normal active regions in PET image segmentation (i.e., relatively high intensity regions compared to background which are considered as neither lesion nor organ). PET organ segmentation is useful when a corresponding computed tomography (CT) image is not available to help with detecting organs. Even if a corresponding CT image is available, usually PET and CT need to be registered. In contrast, for some other applications false negatives should be penalized more, e.g., in ultrasound image segmentation where the boundaries of organs are not very clear, target regions might be under-segmented or in magnetic resonance imaging (MRI) segmentation, small spaces of unsegmented regions within a segmented area might be produced. However, conventional loss functions lack a systematic way of controlling the trade-off between false positive and false negative rates.

Dealing with class-imbalance in image segmentation is critical when training deep models. Designing a proper loss function is a way of handling the imbalance [9]. U-Net (both 2D and 3D) and 2D SegNet minimize cross entropy loss to mimic ground truth segmentation masks for an input image while 3D V-Net applies a Dice based loss function.

Cross entropy is commonly used as a loss function in deep learning. Although it can potentially control output imbalance i.e., false positives and false negatives, it has sub-optimal performance when segmenting highly input class-imbalanced images [10]. There are several ways of handling input imbalance in general classification tasks, e.g., random over/under sampling, synthetic minority over-sampling technique (SMOTE) [11]. Similar to SMOTE, the threshold calibration method, introduced by Pozzolo et al. [12], operates at the data-level, i.e., it requires the data to be under-sampled first. However, this cannot be used when the input is an image and we deal with classifying pixels/voxels (i.e., segmentation). To be specific, it would be meaningless to undersample an image by removing only some of its majority class (e.g., background) pixels/voxels in the case of using full-volumes. Although in patch based approaches, patches can be selected in way to handle the imbalance during training, they do not encode full contextual information and the choice of patch size is not straightforward. Therefore, several different loss-function based techniques such as weighted cross entropy [13], median frequency balancing as used in 2D SegNet [10], the Dice optimization function as used in the 3D V-Net method [9], and a focal loss function [14] have been proposed.

Among all methods introduced for tackling the input-imbalance problem, the Dice based loss function has shown better performance for binary-class segmentation problems [13]. However, the

Table 1: Applied loss functions in the existing deep models for handling imbalance; In-Imb and Out-Imb refer to input imbalance and output imbalance, respectively

Method	Loss function	In-Imb	Out-Imb
2D U-Net [7]	Weighted cross entropy		✓
3D U-Net [8]	Weighted cross entropy		✓
3D V-net [9]	Dice	✓	
Brosch et. al [18]	Sensitivity + specificity	✓	
Sudre et. al [13]	Generalized/weighted Dice	✓	
Berman et. al [19]	Jaccard/IoU	✓	
Lin et. al [14]	Focal		✓
Proposed	Combo	✓	✓

ability of the Dice loss function to control the trade-off between false positives and false negatives (i.e., output imbalance) has not been explored in previous works. Controlling the trade-off is not a trivial issue for some types of medical images and it is not easily handled by a classical Dice optimization function.

Table 1, lists previous works that used different loss functions to cope with input/output imbalance. As reported in the table, none of the current loss functions are able to explicitly handle both input and output imbalance. Some other works attempted to enhance output imbalance in the segmented images using post processing techniques, e.g., Hu et al., [15] applied an energy based refinement step to improve the CNN segmentation results. Similarly, Gibson et al., [16] applied a threshold based refinement step to cope with false positives produced by their convolutional neural network (CNN) based organ segmentation model. Yang et al. [17] also applied a post processing step to reduce both false positives and false negatives in segmented images. In this paper, we leverage both the cross entropy and the Dice optimization functions to define a new loss function that handles both of the aforementioned input and output imbalance types by using global spatial information driven by Dice term and explicitly and gradually enforcing the trade-off between FNs and FPs by cross entropy term.

In this paper, we make the following contributions: a) We introduce a curriculum learning based loss function to handle input and output imbalance (in algorithmic-level) in segmentation problems. b) Our proposed loss improves previous deep models namely 3D U-Net, 3D V-Net,

and our extended version of 2D SegNet i.e., 3D SegNet in both training and testing accuracy for single and multi-organ segmentation from different modalities. c) The proposed loss function, by controlling the trade-off between false positives and negatives, is able to handle missing organs i.e., by penalizing the false positives more. d) We extend 3D U-Net and 3D V-Net from binary to multiclass segmentation models. e) We introduce the first deep volumetric multi-organ semantic model which simultaneously segments and classifies multiple organs from whole body 3D PET scans.

2. Method

Given a medical image volume, the goal is to predict the class of each voxel by assigning an activation value $p(x) \in [0, 1]$ to each voxel x. We adopt a deep learning technique to learn a prediction model $\Phi(x; \theta) : x \rightarrow p(x)$, where θ denotes the model parameters and p_i is activation value for organ/class i.

Cross Entropy Loss Function. For multi-class problems, the cross entropy loss can be computed as $C = \sum_x \sum_i \; t_i \; \ln(p_{i(x)})$ where p is the predicted probability mass function (PMF), which assigns a probability/activation value to each class for each voxel, and t is the one-hot encoded target (or ground truth) PMF, where the index i iterates over the number of organs and x over the number of the samples (i.e., voxels). C can be computed as a sum of several binary cross entropy terms, which for some multi-class problems, as in this paper, makes it possible to have control over false positives/negatives. In the case of binary classification, C can be rewritten as

$$C = \sum_x \; t_i \; \ln(p_i) + (1 - t_i) \; \ln(1 - p_i) \tag{1}$$

The term $(1 - t_i) \; \ln(1 - p_i)$ penalizes false positives as it is zero when the prediction is correct. The binary formulation can also be extended and used for multi-class problems as $\frac{1}{N} \sum_{i=1}^{N} t_i \; \ln(p_i) + (1 - t_i) \; \ln(1 - p_i)$ where $N = number \; of \; classes \times \; number \; of samples$. Therefore, the output is an average of multiple binary cross entropies.

Dice Optimization Function. The Dice function is a widely used metric for evaluating image segmentation accuracy, which can be written in forms of $Dice = True \, positives \; / \; (number \; of \; positives + number \; of \; false \; positives)$ or $Dice = 2 \times TP/(FN + (2 \times TP) + FP)$. It can also be rewritten as a weighted function to generalize into multi-class problems [13]. Recently generalized version of

Dice was introduced [13]. In the binary case, the generalized Dice loss function is written as

$$GDL = 1\left(2\left(\sum_{l=1}^{2} wl \sum_{n} r_{ln} p_{ln}\right) / \left(\sum_{l=1}^{2} wl \sum_{n} r_{ln} + p_{ln}\right)\right) \tag{2}$$

where R and P are the reference foreground segmentation with voxel values r_n and predicted segmentation with voxel values p_n. However, the GDL formulation requires the whole volume to produce meaningful weights (i.e., wl), but in most cases, because of limited GPU memory, the segmentation should be performed on sub-volumes. It is also possible to use weighted version of Dice also known as F_β score [20] to control the trade-off between precision and recall. However, in case of using Dice (F1 or its weighted version F_β) or GDL with sigmoid activation function in output layer of the network to model probabilities, the derivative of the loss in the back-propagation with respect to a specific weight w_{jk} in layer L looks like:

$$\frac{\partial Dice}{\partial w_{jk}^{L}} = \frac{1}{n} \sum_{x} a_k^{L-1} \left(Dice\left(a_j^L, y\right)\right)' \sigma'\left(z_j^L\right) \tag{3}$$

where $\sigma'\left(z_j^L\right)$ is the derivative of the sigmoid activation function. When a neuron has a value close to 0 or 1, the gradient of the sigmoid is very small. As a result, the gradient of the whole cost function with respect to w_{jk} will become very small. Such a saturated neuron will change its weights very slowly. Note that in equation above Dice (F1) can be replaced by F_β or GDL. However, in case of using cross entropy the gradient is computed as

$$\frac{\partial CE}{\partial w_{jk}^{L}} = \frac{1}{n} \sum_{x} a_k^{L-1} \left(\sigma\left(z_j^L\right) - y\right) \tag{4}$$

Here gradient is not affected by $\sigma'\left(z_j^L\right)$ anymore, so the gradient only depends on the neurons output, the target y and the neurons input a_k^{L-1}. This avoids learning slow-down and helps with the vanishing gradient problem from which deep neural networks suffer.

Combo Loss. To leverage the Dice function that handles the input class-imbalance problem, i.e., segmenting a small foreground from a large context/background, while at the same time controlling the trade-off between FP and FN and enforcing a smooth training using cross entropy as discussed above, we introduce our loss L as a weighted sum of two terms: A Dice loss and a modified cross entropy (i.e. adding β and $1-\beta$ to control output FP and FN) to encode curriculum learning, and is written as:

$$L = \alpha\left(-\frac{1}{N}\sum_{i=1}^{N}\beta\left(t_i - \ln p_i\right) + (1-\beta)\left[(1-t_i)\ln\left(1-p_i\right)\right]\right) - (1-\alpha)\sum_{i=1}^{K}\left(\frac{2\sum_{i=1}^{N} p_i t_i + S}{\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} t_i + S}\right) \quad (5)$$

where α controls the amount of Dice term contribution in the loss function L, and $\beta \in [0, 1]$ controls the level of model penalization for false positives/negatives: when β is set to a value smaller than 0.5, FP are penalized more than FN as the term $(1 - t_i)\ \ln\ (1 - p_i)$ is weighted more heavily, and vice versa. In our implementation, to prevent division by zero, we perform add-one smoothing (a specific instance of the additive/Laplace/Lidstone smoothing) [21], i.e., we add unity constant S to both the denominator and numerator of the Dice term. Although the proposed loss seems to be simply combining two different loss functions, we deliberately chose the binary version of the cross entropy to enable us to explicitly enforce a the *intended* trade-off between false positives and negatives using the parameter β (equation 5) and, at the same time, keep the model parameters out of bad local minima via the global spatial information provided by Dice term. In L, curriculum learning can be seen as a strategy for improving the convergence of highly nonconvex optimization problems [22]. In the initial stages of the model optimization of our approach, if the Dice term acts alone it may tend to converge to an overfit solution quickly. However, the addition of the cross-entropy term acts as a regularizer for the Dice term and avoids large model parameter changes that may lead to unsuitable model parameters.

After sigmoid normalization over all the channels (i.e., classes) in the last layer, the Combo loss function is computed using the flattened volumes (one-hot multi-label encoding for both the predicted and ground truth volumes containing several objects) of size $W \times H \times D \times C$ where W, H, D, and C refer to width, height, depth, and number of channelsclasses. This strategy makes it simple to generalize to multi-class segmentation hence directly controlling FPs and FNs over entire volume.

Model Parameter Optimization. To optimize the model parameters θ to minimize the loss, we use error back propagation, which relies on the chain rule. We calculate the gradient of L with respect to p_j, i.e., dL/dp_j,

Table 2: Comparison between our proposed method and state of the art deep learning segmentation methods. Up_S refer to up-sampling type.

Network	SC	Loss	Up_S	Params
2D SegNet [10]	No	Cross entropy	Specific	16,375,169
3D U-Net [8]	Yes	Cross entropy	Regular	12,226,243
3D V-Net [9]	Yes	Dice	Regular	84,938,241
Proposed	No	Proposed	Regular	13,997,827

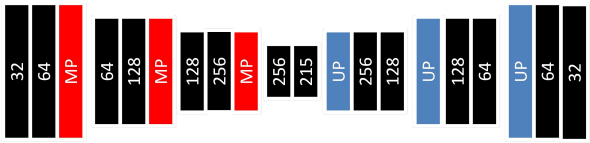

Figure 1: The applied network architecture. Black: convolution $3 \times 3 \times 3$, MP: max-pooling $2 \times 2 \times 2$, UP: up-sampling $2 \times 2 \times 2$. The values inside the boxes show number of channels.

$$\frac{\partial L}{\partial p_j} = 2\alpha \times \left(-\frac{1}{N} \sum_{i=1}^{N} \beta \left(\frac{t_i}{p_i} \right) + (1-\beta) \left(-\frac{1-t_i}{1-p_i} \right) \right) - (1-\alpha) \sum_{i=1}^{N} \frac{t_i \left(\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} t_i \right) - p_j \left(\sum_{i=1}^{N} p_i t_i \right)}{\left(\sum_{i=1}^{N} p_i + \sum_{i=1}^{N} t_i \right)^2} \quad (6)$$

Then we calculate how the changes in the model parameters in the last layer of the deep architecture affect the predicted p_j, and so on.

Deep Model Architecture. We use the deep architecture shown in Fig. 1. This architecture departs from existing architectures like 3D U-Net, 3D U-Net, and 2D SegNet as listed in Table 2. We adopt this simple network to show that the improvement in results is not attributed to some elaborate architecture and to validate our hypothesis that, even with a simple shallower architecture as long as a proper loss function is used, it is possible outperform more complex architectures e.g., networks with skip connections [7–9] or specific up-sampling [10].

Training. For multi-organ segmentation from whole body PET images, as the volumes are too large to fit into memory, we extract random sub-volumes from each whole body scan to train a model. Each sub-volume could include voxels belonging to $n \in \{0, 1, ..., K\}$ organs, with $n = 0$

indicating a sub-volume including only background. However, for binary segmentation i.e., 3D ultrasound and MRI datasets, we train using the entire volumes. On test data (only for PET), we apply a volumetric sliding window (with stride), i.e., a volumetric field of view V is partitioned into smaller sub-volumes $\{v_1, v_2, ...v_n\}$, where the size of v_i is the same as that of the training sub-volumes. Along any of the dimensions, the stride would be at least 1 voxel and at most the size of the sub-volume in that dimension. Larger strides speed up the computation at the expense of coarser spatial predictions. Let v_i be a subvolume with activation a_{v_i}, V_x be the set of subvolumes that include x, A_{V_x} be the set of corresponding activation values. $T(A_{V_x})$ is the set of indicator variables whose value is 1 if the activation is larger than t, and 0 otherwise, where t is a threshold value. Then, the the label assigned to voxel x is given by: $f(x) = \begin{cases} 0, & \text{if} \quad \max(T(A_{V_x})) = 0 \\ 1, & \text{otherwise} \end{cases}$. In other words, a single voxel x may reside within multiple overlapping subvolumes; if the activation of any these subvolumes is larger than threshold T, then x is assigned 1, and 0 otherwise.

3. Implementation details

a) PET multi-organ segmentation: For training the PET multi-organ segmentation network, from each training image, we extract 100 randomly positioned $80 \times 80 \times 80$-voxel sub-volumes per organ (5 organs in total: brain, heart, left kidney, right kidney, and bladder) and another 100 for negative background sub-volumes. Therefore, we train all the models with $\sim 600\times$ number of training volumes. In test, the striding size was set to $20\times20\times20$. PET volumes size varied from $128\times128\times \sim 200$ to $128 \times 128\times \sim 500$. We train and test all the models using two Titan-X GPUs in parallel each with batch-size 1.

b) Ultrasound echocardiography and prostate MRI segmnetation: We train and test all the models on these datasets with whole-volume images (i.e., not sub-volumes) of size $128 \times 128 \times 128$ and $48 \times 256 \times 256$ for ultarsound and MRI datasets, respectively using two M5000 GPUs in parallel each with batch-size 2. As explored by Masters et al., [23], small mini-batch sizes can provide more up-to-date gradient calculations, which results in more stable and reliable training while reducing over-fitting more compared to larger batch sizes.

As MRI and ultrasound images are taken from a part of the body, the number of slices per volume are relatively less compared to PET whole body volumes. To prevent sliding window for both training and testing and fitting whole MRI and ultrasound volumes into memory, we slightly

resampled MRI and Ultrasound images without losing much information causes by resampling. However, PET volumes should be highly resized in order to be fitted into memory which results in considerable accuracy drop. Therefore, we did not resample PET images.

For all datasets, we initialize our models and competing methods using the method introduced by and Bengio [24] and train them with ADADELTA [25], with learning rate of 1, $\rho = 0.95$, $\epsilon = 1e - 08$, and $decay = 0$. All the models are coded using TensorFlow. To prevent gradient vanishing/exploding we use batch normalization after each convolution layer [26]. It also allows us to use higher learning rate. Similar to how hyperparameters values are selected in deep models, e.g., learning rate and pooling window size, the optimal values for α and β were also found by grid search to optimize results on the validation set (i.e., one round of cross-validation). We found that the equal contribution (i.e., $\alpha = 0.5$) of Dice and cross-entropy terms gives the best results. We experimentally found that different data can be sensitive to either FP or FN or both e.g. PET suffers more from false positive, due to noise and to small metabolic activities all over the body ($\beta = 0.4$). Ultrasound and MRI data, on the other hand, suffer more from under segmentation i.e. false negatives ($\beta = 0.6$ for MRI and $\beta = 0.7$ for ultrasound images), which is due to unclear borders of left ventricle and prostate, respectively. Therefore we set different values to handle the imbalance properly for each data. For the last layer of the proposed method, we applied the sigmoid activation function as it allowed us to compute the loss over only foreground objects (i.e., there is no extra channel for the background class, as softmax function requires) and then normalize the output into the range [0-1]. To obtain the segmentation masks we use threshold of 0.5.

All the models have been trained for a fixed number of epochs and we report the results for the best epoch based on the validation set. Note that for the competing methods we set the hyper-parameters as proposed by the authors of these methods. For fairness and to elucidate the direct effect of the proposed Combo loss, when we replace the original loss functions of the competing methods with Combo loss (Tables 3 and 5), we do not change the original network hyper-parameters.

4. Datasets

For evaluation, we use three different datasets: a) 58 whole body PET scans of resolution $\sim 0.35\times \sim 0.35\times \sim 3.5 - 5\ mm$. We randomly pick 10 whole body volumes for testing and train with the 48 remaining volumes. We normalize the intensity range of our training and testing volumes using the min-max method based on min and max intensity values of the whole training

set. Next, in both training and testing, each single sub-volume is also normalized to $[0, 1]$ using its min and max before feeding it into network. b) 958 MRI prostate scans of different resolution which were resampled to voxel size of $1 \times 1 \times 3$ (mm). We randomly picked 258 volumes for testing, and train with remaining 700 volumes. c) hlUltrasound echocardiography images of resolution $2 \times 2 \times 2$ (mm), used for left ventricular myocardial segmentation, were split into 430 train and 20 test. The datasets were collected internally and from The Cancer Imaging Archive (TCIA) QIN-HEADNECK and ProstateX datasets [27–31]. Samples of the three datasets are shown in Fig. 2. The ground truth for ultrasound and MRI datasets were generated by experts in medical imaging and anatomy using the established Siemens annotation pipeline which has been adopted previously in a large number of research projects and published works by Siemens and collaborators., The PET ground truth segmentation has been completed by a medical image analysis researcher with several years of experience.

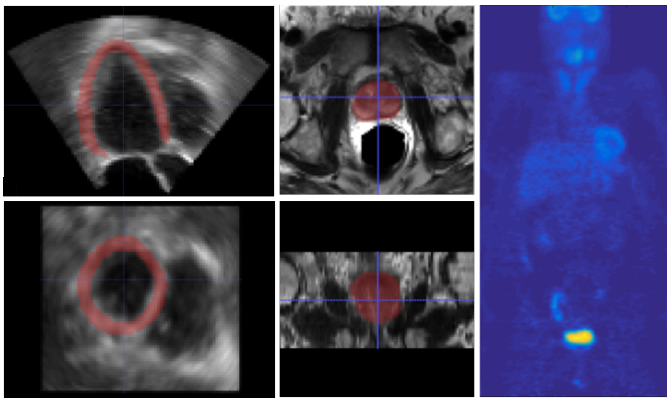

Figure 2: Samples of the used datasets. The first column shows the left ventricular myocardium (highlighted in red) in ultrasound echocardiography. In the second column, the prostate is highlighted in red in an MRI scan. For MRI and ultrasound samples, the 2 rows show the axial and coronal views. Column 3 shows coronal view of a whole body PET scan.

5. Results

We perform a set of experiments on three different datasets to show the effectiveness of the proposed method on both input and output imbalance. Note that segmentation models suffer less from input imbalance when organs are not relatively much smaller than other organs, but suffer more from output imbalance as the boundaries of the organs are clearly visible, e.g. in ultrasound

Table 3: Comparing the performance of competing methods with/without the proposed loss function vs. the proposed method for PET multi-organ segmentation

	Methods	Jaccard	Dice	FPR	FNR
a	3D U-Net [8]	0.55 ± 0.16	0.69 ± 0.16	0.41 ± 0.31	0.30 ± 0.15
	3D U-Net_Combo	0.68 ± 0.18	0.79 ± 0.15	0.18 ± 0.09	0.21 ± 0.17
b	3D V-Net [9]	0.52 ± 0.17	0.67 ± 0.16	0.89 ± 0.90	0.13 ± 0.09
	3D V-Net_Combo	0.55 ± 0.17	0.70 ± 0.15	0.64 ± 0.46	0.16 ± 0.11
c	3D SegNet	0.41 ± 0.20	0.56 ± 0.21	0.66 ± 0.78	0.32 ± 0.28
	3D SegNet_Combo	0.55 ± 0.19	0.69 ± 0.17	0.37 ± 0.38	0.28 ± 0.17
d	Ahmadvand et al. [32]	0.41 ± 0.18	0.53 ± 0.23	0.35 ± 0.77	0.37 ± 0.82
e	P_{CE}	0.34 ± 0.15	0.49 ± 0.18	0.67 ± 0.42	0.48 ± 0.14
	P_{WCE}	0.45 ± 0.20	0.60 ± 0.19	0.46 ± 0.23	0.37 ± 0.22
f	P_{D}	0.67 ± 0.09	0.80 ± 0.07	0.43 ± 0.19	0.06 ± 0.04
	P_{DCE}	0.73 ± 0.10	0.84 ± 0.07	0.09 ± 0.06	0.21 ± 0.11
	$\mathbf{P_{Combo}}$	0.73 ± 0.13	0.84 ± 0.10	0.07 ± 0.02	0.22 ± 0.14

or normal metabolic activity in PET might cause false positives. Our evaluation is divided into 2 parts. First, in subsection 5.1, we compare, both qualitatively and quantitatively, the performance of all the competing methods to the proposed method on the test data, for multi-organ segmentation from PET scans. We test different modification/variants of the proposed loss with the proposed architecture, i.e., cross entropy optimization (P_{CE}, weighted cross entropy (P_{WCE}), Dice optimization (P_{D}), Dice + cross entropy optimization (P_{DCE}), and the proposed loss (P_{Combo}). DCE refers to simply integrating Dice and traditional cross entropy losses, whereas, Combo refers to combining the weighted version of cross entropy with Dice. Second, in subsection 5.2, we perform similar experiments to subsection 5.1 for single organ segmentation from two more different modalities, i.e., MRI and ultrasound scans.

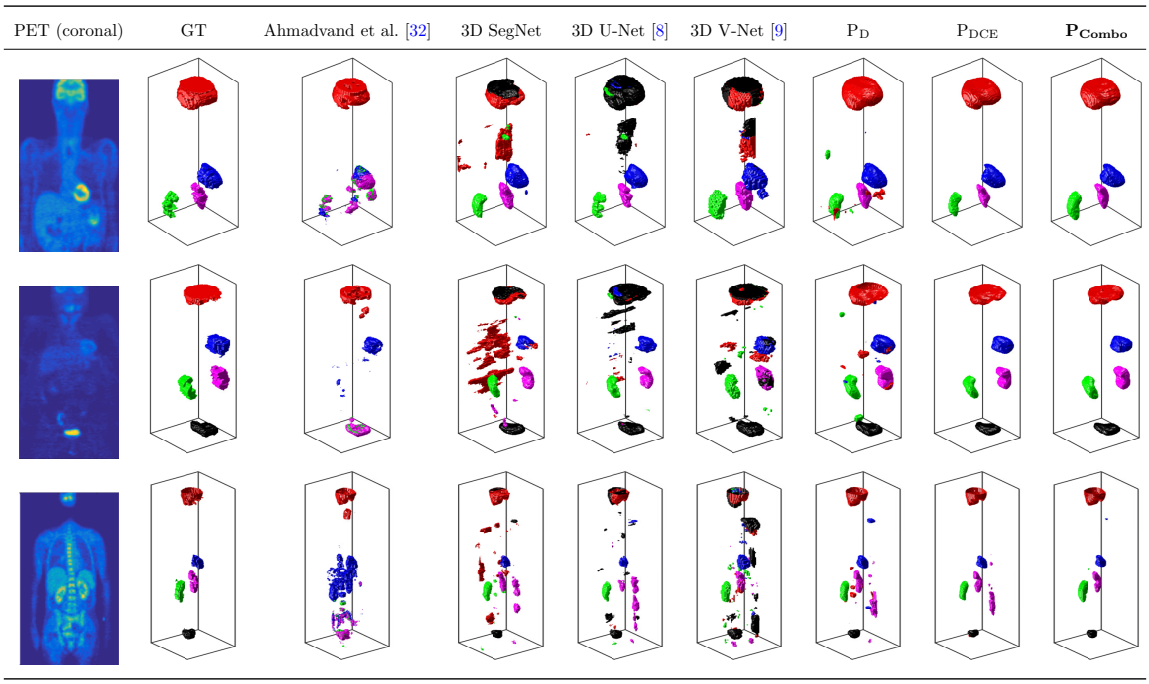

Figure 3: Comparing multi-organ localization-segmentation-classification results of the proposed vs. competing methods. Each row shows a sample patient data.

5.1. Performance of the proposed vs. competing methods on multi-organ PET segmentation

We treat the multi-class case as binary, i.e., the one-hot multi-label encoding for the both the predicted and ground truth volumes containing several objects are flattened and the Combo loss is computed. In this case, similar to binary segmentation, balancing the false positives and negatives improves segmentation. As reported in Table 3, the proposed architecture with proposed loss (P_{Combo}) outperforms all competing methods with $57\% \pm 24\%$, $38\% \pm 18\%$, $86\% \pm 5\%$ in Jaccard, Dice and FPR, respectively. Comparing rows of section a in Table 3, we note that: Modified 3D U-Net improves with our proposed loss (Combo) relatively by 23.6%, 14.5%, 56%, and 30% in Jaccard, Dice, FPR, and FNR, respectively. Comparing rows of section b, we note that: 3D V-Net improves with our proposed loss relatively by 5.8%, 4.5%, and 28%, in Jaccard, Dice, and FPR, respectively. Section c shows that 3D SegNet improves with our proposed loss by relatively 34.1%, 23.2%, 44%, and 12.5% in Jaccard, Dice, FPR, and FNR, respectively. Comparing $\mathrm{P_{CE}}$ vs. $\mathrm{P_{WCE}}$ in section e of Table 3 shows that WCE helps. Comparing $\mathrm{P_{D}}$ vs. $\mathrm{P_{Combo}}$ shows that the proposed Combo loss improves the results. Although the results and formulation of Dice + original cross entropy (i.e., DCE) and Combo loss are close, it is important to note that, in the Combo loss formulation, we weight the two terms of the original cross entropy so we can enforce the intended trade-off between FP and FN.

As shown in Figure 3, although 3D U-Net, 3D V-Net, and the extended version (3D) of SegNet are able to locate the normal activities (bright areas in the image because of absorbing radio-tracer. The look very similar to abnormalities) and segment them, two issues are visible: a) misclassification of organs: the competing methods were not successful in distinguishing the organs from each other, as sometimes the brain (red) has been labeled as bladder (black); b) the competing methods tend to produce false positives i.e., wrongly labeling some background voxels as an organ (or one organ as another) or missing an organ (false negative). As shown in the figure, $\mathrm{P_{D}}$ still produces false positives, but no misclassification of organs. $\mathrm{P_{CE}}$ shows clearer segmentations, however, as we penalize the false positives more with the proposed loss we obtain much clearer outputs (last columns: $\mathrm{P_{Combo}}$). The performance of the proposed method was evaluated for each specific organ and reported in Table 4.

Over all the organs, Dice scores for the proposed method (proposed architecture + Combo loss) ranges from 0.58 to 0.91. We show the worst, an in-between and the best results in terms of Dice score in Fig. 5. Although the left case in the figure seems to be the worst result in terms of Dice

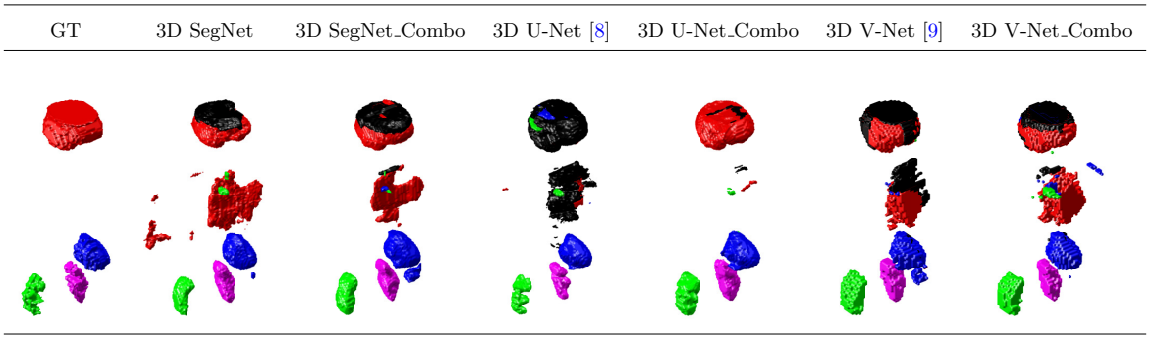

Figure 4: Competing methods' results before and after adding proposed loss function.

Table 4: Organ-specific quantitative results of the proposed method for PET dataset. BR, HR, LK, RK, and BL refer to different organs i.e., brain, heart, left kidney, right kidney, and bladder, respectively. The numbers in parentheses in front of each organ show the average percentage number of voxels which belong to that organ in whole volumes.

	Jaccard	Dice	FPR	FNR
BR ($\sim 0.64\%$)	0.74 ± 0.20	0.83 ± 0.16	0.06 ± 0.04	0.21 ± 0.22
HR ($\sim 0.14\%$)	0.65 ± 0.15	0.78 ± 0.12	0.11 ± 0.07	0.29 ± 0.15
LK ($\sim 0.08\%$)	0.68 ± 0.08	0.81 ± 0.06	0.19 ± 0.15	0.20 ± 0.07
RK ($\sim 0.09\%$)	0.68 ± 0.10	0.81 ± 0.07	0.09 ± 0.07	0.26 ± 0.11
BL ($\sim 0.09\%$)	0.69 ± 0.12	0.81 ± 0.09	0.04 ± 0.05	0.28 ± 0.14

score, it is a difficult case with several missing organs. However, the proposed method has been able to handle multiple missing organs to a high extent. Note that some organs c an be physically absent from a patient body, as in renal agenesis or radical (complete) nephrectomy, but in PET scans, there might be more "missing" organs (similar to the left case in Fig. 5) simply because of lack of radiotracer uptake in these organs thus they do not appear in PET. Although, in training, Dice score improvement compared to 3D V-Net is small, as shown in Figure 4, in test, proposed loss helped 3D V-Net in terms of reducing organ misclassification and false positives. Looking at both Table 3 and Fig. 4, 3D U-Net and 3D SegNet achieved higher performance when incorporating the proposed loss.

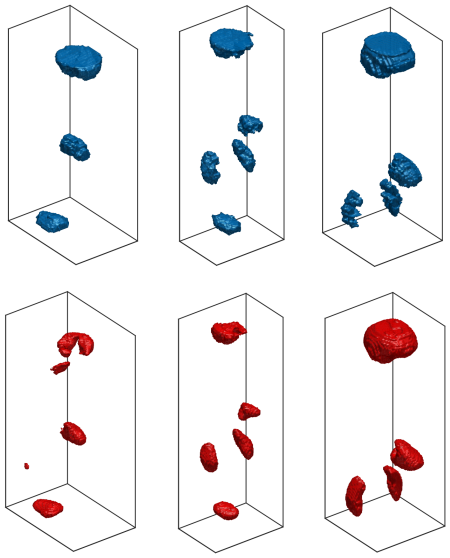

Figure 5: Sample segmentations by the proposed method. First row shows ground truth segmentations and second row shows the proposed method results. From left to right: worst (Dice = 0.58), in-between (Dice = 0.76), and best (Dice = 0.91)

5.2. Performance of the proposed vs. competing methods on single organ segmentation from MRI and ultrasound

For MRI and ultrasound datasets, we observed that all the methods are more prone to false negatives than false positives, so we weigh more the false negative term of the proposed loss (i.e., increase β to 0.9). As reported in Table 5, similar to results in Section 5.1, the Combo loss function improved 3D U-Net and 3D V-Net by 4.6% and 1.13% in Dice and 43.8% and 16.7%in FNR, respectively, for MRI prostate segmentation. Similarly, 3D U-Net and 3D V-Net results were improved by 8.23% and 3.4% in Dice and 33.3% and 16.7%in FNR, respectively, for ultrasound left ventricular myocardial segmentation.

As can also be seen in Table 5, the proposed loss also helps reduce the variance of the segmentation results.

We also compared the proposed loss function with the recently introduced Focal loss function [14]. Our integrative loss function outperformed Focal loss after both were applied to different networks (Table 5). We applied Focal loss to the best performing competing method for each dataset i.e., 3D V-Net for MRI and 3D U-Net for ultrasound dataset. For Focal loss, we tested

Table 5: Comparing performance of competing methods with/without proposed (_Combo) loss function vs. proposed method for MRI prostate and ultrasound left ventricular myocardial segmentation. The average percentage of voxels belonging to organ in whole volumes are 1.01% and 0.6% for left ventricle and prostate in ultrasound and MRI volumes, respectively.

	Methods	Dice	FPR	FNR
MRI	3D U-Net [8]	0.87 ± 0.07	0.0004 ± 0.0004	0.16 ± 0.12
	3D U-Net_Combo	0.91 ± 0.05	0.0005 ± 0.0005	0.09 ± 0.08
	3D V-Net [9]	0.88 ± 0.05	0.0006 ± 0.0004	0.12 ± 0.08
	3D V-Net_Focal	0.87 ± 0.04	0.0002 ± 0.0002	0.19 ± 0.07
	3D V-Net_Combo	0.89 ± 0.05	0.0006 ± 0.0005	0.10 ± 0.08
	ProposedArc_Combo	0.90 ± 0.04	0.0007 ± 0.0006	0.08 ± 0.07
Ultrasound	3D U-Net [8]	0.85 ± 0.05	0.0020 ± 0.0006	0.12 ± 0.12
	3D U-Net_Combo	0.92 ± 0.05	0.0007 ± 0.0003	0.08 ± 0.09
	3D U-Net_Focal	0.88 ± 0.11	0.0004 ± 0.0005	0.17 ± 0.15
	3D V-Net [9]	0.84 ± 0.04	0.0020 ± 0.0008	0.12 ± 0.08
	3D V-Net_Combo	0.87 ± 0.03	0.0020 ± 0.0005	0.10 ± 0.04
	ProposedArc_Combo	0.92 ± 0.05	0.0006 ± 0.0004	0.09 ± 0.10

several different values for α and γ, but as suggested by the authors we obtained better results with $\alpha = 0.25$ and $\gamma = 2.0$. Note that there is no correspondence between the alpha used in the Focal loss paper (the weight assigned to the rare class) and the one we use in Combo loss equation (the weight that controls the contribution of Dice and cross entropy terms). For the MRI dataset, the proposed Combo loss outperformed Focal loss by 2.3% and 47.4% in Dice and FNR, respectively, when both were used in 3D V-Net. For the ultrasound dataset, Combo loss outperformed Focal loss by 10.8% in Dice. In Figure 6, we plot both Dice and Hausdorff distance (HD) of the Combo loss vs. competing methods. As shown in the figure, the proposed method outperforms the competing methods in terms of Dice score. Comparing both Dice and Hausdorff distance values of the competing methods, after applying Combo loss (i.e., U_C, and V_C) in Figure 6, the range of the values are smaller, i.e., less outliers compared to when they use original loss (i.e., U and V).

Among the competing methods, U-Net applies cross entropy loss while V-Net leverages Dice loss. To show the direct contribution of the Combo loss, we replace the original loss functions in U-Net and V-Net with Combo (Table 5). As reported in Table 5, after replacing cross entropy loss of U-Net with Combo loss, the Dice scores improve from 0.87 to 0.91 and 0.85 to 0.92 for MRI and ultrasound datasets, respectively. Similarly, when replacing the Dice loss function of V-Net with the proposed Combo loss, the segmentation results improve from 0.88 to 0.89 and 0.84 to 0.87 for MRI and ultrasound datasets, respectively.

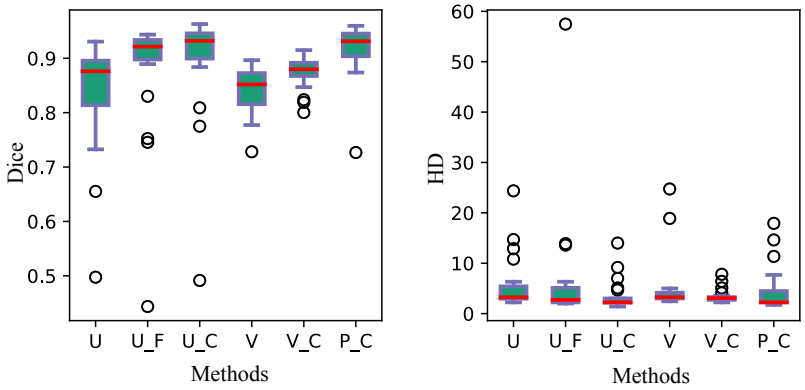

Figure 6: Dice and Hausdorff distance (HD) of the competing methods vs. the proposed Combo loss for the ultrasound dataset. In the figure, U, U_F, U_C, V, V_C, and P_C refer to 3D U-Net, 3D U-Net_Focal, 3D U-Net_Cmbo, 3D V-Net, 3D V-Net_Combo, and ProposedArc_Combo, respectively.

Parameter α controls the contribution of Dice and cross entropy terms while parameter β in

the second term, i.e., cross entropy, controls the trade-off between false positives and negatives. As a key contribution of the paper is providing the means to explicitly control output balance, i.e., false positives and negatives, we tested several different values for parameter beta to see how the final results are affected by β and we fix parameter α that controls the trade-off between Dice and cross entropy to 0.5. In Figure 7, we show the different Dice and HD results obtained from different β values, which control false positives and false negatives. As expected, we note that the final segmentations are affected by the choice of parameter beta and the best results in terms of higher Dice and lower Hausdorff distance were obtained for $\beta = 0.7$ and $\beta = 0.6$ for ultrasound and MRI datasets, respectively. As HD is sensitive to outliers, there are sometimes relatively large values in the HD results (i.e., second column in the figure)

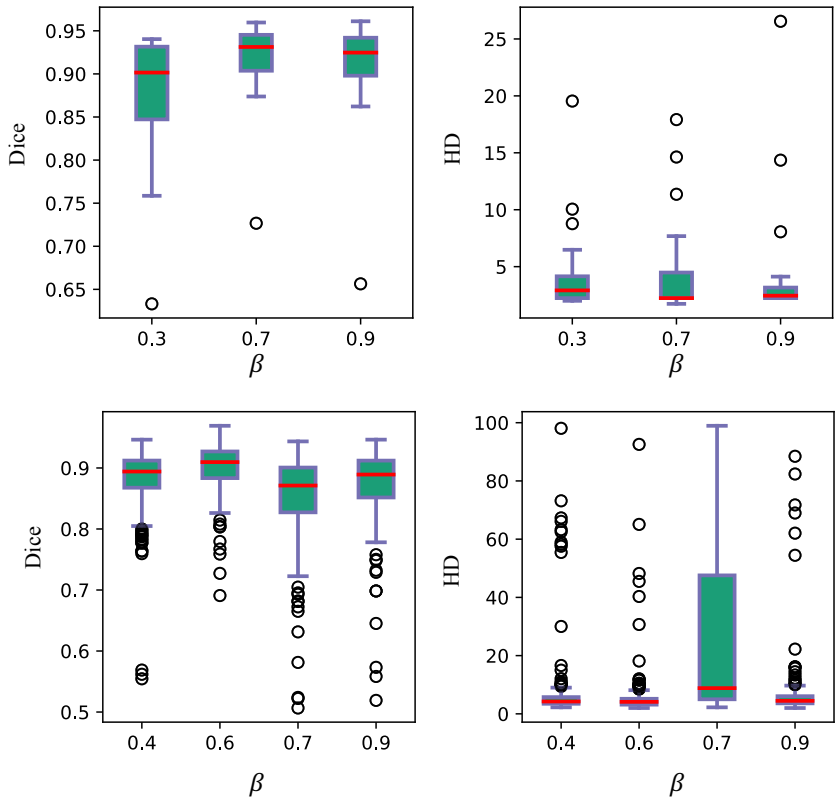

Figure 7: Dice and Hausdorff distance (HD) with various parameter β values. First and second rows show ultrasound and MRI results, respectively.

6. Conclusion

In this paper, we proposed a curriculum learning based loss function to handle input/class-imbalance and output imbalance (i.e., enforcing the trade-off between false positives and false negatives). Note that enforcing a desired trade-off between false positives and false negatives can be seen in Tables 3 and 5). Noting the change in FPR and FNR values of 3D U-Net, 3D V-Net, and 3D Seg-Net when they apply Combo loss, we see that FPR or FNR is severely decreased when the models are penalized for FP or FN, respectively (for PET data i.e., Table III, the Combo loss penalizes FP and for MRI and ultrasound data i.e., Table V, it penalizes FN). The proposed loss function resulted in improved performance in both multi- and single-organ segmentation from different modalities. The proposed loss function also improved the existing methods in terms of achieving higher Dice and lower false positive and false negative rates. In this work, we applied the proposed loss function to a multi-organ segmentation problem, but it can simply be leveraged for other segmentation tasks as well. The key advantage of the proposed Combo loss is that it enforces a *desired* trade-off between the false positives and negatives (which results in cutting out post-processing) and avoids getting stuck in bad local minima as it leverages Dice term. The Combo loss converges considerably faster than cross entropy loss during training. Similar to Focal loss, our Combo loss also has two parameters that need to be set. In this work, we used cross-validation to set the hyperparameters (including α and β of our proposed loss). Future work can explore using Bayesian approaches [33, 34].

References

[1] Y. Yuan, M. Chao, Y.-C. Lo, Automatic skin lesion segmentation using deep fully convolutional networks with jaccard distance, IEEE Transactions on Medical Imaging 36 (9) (2017) 1876–1886.

[2] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. 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.

[3] C. F. Baumgartner, L. M. Koch, M. Pollefeys, E. Konukoglu, An exploration of 2D and 3D deep learning techniques for cardiac MR image segmentation, arXiv preprint arXiv:1709.04496.

[4] F. Milletari, S.-A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Bötzel, et al., Hough-CNN: deep learning for segmentation of deep brain regions in mri and ultrasound, Computer Vision and Image Understanding 164 (2017) 92–102.

[5] G. Wang, M. A. Zuluaga, W. Li, R. Pratt, P. A. Patel, M. Aertsen, T. Doel, A. L. David, J. Deprest, S. Ourselin, et al., Deepigeos: a deep interactive geodesic framework for medical image segmentation, arXiv preprint arXiv:1707.00652.

[6] A. BenTaieb, G. Hamarneh, Topology aware fully convolutional networks for histology gland segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2016, pp. 460–468.

[7] 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 (MICCAI), Springer, 2015, pp. 234–241.

[8] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, O. Ronneberger, 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2016, pp. 424–432.

[9] F. Milletari, N. Navab, 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.

[10] V. Badrinarayanan, A. Kendall, R. Cipolla, SegNet: a deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (12) (2017) 2481–2495.

[11] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, Smote: synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[12] A. Dal Pozzolo, O. Caelen, R. A. Johnson, G. Bontempi, Calibrating probability with under-sampling for unbalanced classification, in: Computational Intelligence, 2015 IEEE Symposium Series on, IEEE, 2015, pp. 159–166.

[13] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, M. J. Cardoso, Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations, in: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, Springer, 2017, pp. 240–248.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, arXiv preprint arXiv:1708.02002.

[15] P. Hu, F. Wu, J. Peng, Y. Bao, F. Chen, D. Kong, Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets, International Journal of Computer Assisted Radiology and Surgery 12 (3) (2017) 399–411.

[16] E. Gibson, F. Giganti, Y. Hu, E. Bonmati, S. Bandula, K. Gurusamy, B. R. Davidson, S. P. Pereira, M. J. Clarkson, D. C. Barratt, Towards image-guided pancreas and biliary endoscopy: Automatic multi-organ segmentation on abdominal CT with dense dilated networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), Springer, 2017, pp. 728–736.

[17] D. Yang, D. Xu, S. K. Zhou, B. Georgescu, M. Chen, S. Grbic, D. Metaxas, D. Comaniciu, Automatic liver segmentation using an adversarial image-to-image network, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2017, pp. 507–515.

[18] T. Brosch, Y. Yoo, L. Y. Tang, D. K. Li, A. Traboulsee, R. Tam, Deep convolutional encoder

networks for multiple sclerosis lesion segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 3–11.

[19] M. B. A. R. T. Matthew, B. Blaschko, The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks.

[20] S. R. Hashemi, S. S. M. Salehi, D. Erdogmus, S. P. Prabhu, S. K. Warfield, A. Gholipour, Asymmetric similarity loss function to balance precision and recall in highly unbalanced deep medical image segmentation. arxiv preprint, arXiv preprint arXiv:1803.11078.

[21] S. J. Russell, P. Norvig, Artificial intelligence: a modern approach, Malaysia; Pearson Education Limited,, 2016.

[22] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 41–48.

[23] D. Masters, C. Luschi, Revisiting small batch training for deep neural networks, arXiv preprint arXiv:1804.07612.

[24] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, 2010, pp. 249–256.

[25] M. D. Zeiler, Adadelta: an adaptive learning rate method, arXiv preprint arXiv:1212.5701.

[26] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, arXiv preprint arXiv:1502.03167.

[27] R. R. Beichel, E. J. Ulrich, C. Bauer, A. Wahle, B. Brown, T. Chang, K. A. Plichta, B. J. Smith, J. J. Sunderland, T. Braun, A. Fedorov, D. Clunie, M. Onken, J. Riesmeier, S. Pieper, R. Kikinis, M. M. Graham, T. L. Casavant, M. Sonka, J. M. Buatti, Data from QIN-HEADNECK http://doi.org/10.7937/k9/tcia.2015.k0f5cgli, The Cancer Imaging Archive.

[28] A. Fedorov, D. Clunie, E. Ulrich, C. Bauer, A. Wahle, B. Brown, M. Onken, J. Riesmeier, S. Pieper, R. Kikinis, J. Buatti, R. Beichel, DICOM for quantitative imaging biomarker development: a standards based approach to sharing clinical data and structured PET/CT analysis results in head and neck cancer research. peerj 4:e2057 https://doi.org/10.7717/peerj.2057.

[29] L. Geert, D. Oscar, B. Jelle, K. Nico, H. Henkjan, Prostatex challenge data. https://doi.org/10.7937/k9tcia.2017.murs5cl, The Cancer Imaging Archive.

[30] G. Litjens, O. Debats, J. Barentsz, N. Karssemeijer, H. Huisman, Computer-aided detection of prostate cancer in MRI, IEEE Transactions on Medical Imaging 33 (5) (2014) 1083–1092.

[31] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, et al., The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository, Journal of Digital Imaging 26 (6) (2013) 1045–1057.

[32] P. Ahmadvand, N. Duggan, F. Bénard, G. Hamarneh, Tumor lesion segmentation from 3D PET using a machine learning driven active surface, in: International Workshop on Machine Learning in Medical Imaging, Springer, 2016, pp. 271–278.

[33] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, in: Advances in neural information processing systems, 2012, pp. 2951–2959.

[34] P. Murugan, Hyperparameters optimization in deep convolutional neural network/bayesian approach with gaussian process prior, arXiv preprint arXiv:1712.07233.