Esophageal Tumor Segmentation in CT Images using a Dilated Dense Attention Unet (DDAUnet)

SAHAR YOUSEFI[1], HESSAM SOKOOTI[1], MOHAMED S. ELMAHDY[1], IRENE M. LIPS[2],
MOHAMMAD T. MANZURI SHALMANI[3], ROEL T. ZINKSTOK[2,4], FRANK J.W.M. DANKERS[2],
AND MARIUS STARING[1,2,5]

[1]Department of Radiology, Leiden University Medical Center, Leiden, The Netherlands.
[2]Department of Radiation Oncology, Leiden University Medical Center, Leiden, The Netherlands.
[3]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran.
[4]Department of Radiation Oncology, Netherlands Cancer Institute, Amsterdam, The Netherlands.
[5]Delft University of Technology, Delft, The Netherlands.

Corresponding author: Sahar Yousefi (e-mail: s.yousefi.radi@lumc.nl).

ABSTRACT Manual or automatic delineation of the esophageal tumor in CT images is known to be very challenging. This is due to the low contrast between the tumor and adjacent tissues, the anatomical variation of the esophagus, as well as the occasional presence of foreign bodies (e.g. feeding tubes). Physicians therefore usually exploit additional knowledge such as endoscopic findings, clinical history, additional imaging modalities like PET scans. Achieving his additional information is time-consuming, while the results are error-prone and might lead to non-deterministic results. In this paper we aim to investigate if and to what extent a simplified clinical workflow based on CT alone, allows one to automatically segment the esophageal tumor with sufficient quality. For this purpose, we present a fully automatic end-to-end esophageal tumor segmentation method based on convolutional neural networks (CNNs). The proposed network, called Dilated Dense Attention Unet (DDAUnet), leverages spatial and channel attention gates in each dense block to selectively concentrate on determinant feature maps and regions. Dilated convolutional layers are used to manage GPU memory and increase the network receptive field. We collected a dataset of 792 scans from 288 distinct patients including varying anatomies with air pockets, feeding tubes and proximal tumors. Repeatability and reproducibility studies were conducted for three distinct splits of training and validation sets. The proposed network achieved a DSC value of 0.79 ± 0.20, a mean surface distance of 5.4 ± 20.2mm and 95% Hausdorff distance of 14.7 ± 25.0mm for 287 test scans, demonstrating promising results with a simplified clinical workflow based on CT alone. Our code is publicly available via https://github.com/yousefis/DenseUnet_Esophagus_Segmentation.

INDEX TERMS Esophageal tumor segmentation, CT images, densely connected pattern, UNet, dilated convolutional layer, attention gate

I. INTRODUCTION

Esophageal cancer is one of the least studied cancers [1], while it is lethal in most patients [2]. Because of the very poor survival rate, three standard treatment options are available, i.e. chemoradiotherapy (CRT), neoadjuvant CRT followed by surgical resection, or radical radiotherapy [3]. For this purpose, rapid and accurate delineation of the target volume in CT images plays a very important role in therapy and disease control. The complexities raised by automatic esophageal tumor delineation in CT images can be divided into several categories: textural similarities and the absence of contrast between the tumor and its adjacent tissues; the anatomical variation of different patients either intrinsically or caused by a disease, like a hiatal hernia in which part of the stomach bulges into the chest cavity through an opening of the diaphragm (see Figure 1-(e) and (i)), extension of tumor into the stomach, or existence of

air pocket inside the esophagus; existence of foreign bodies during the treatment, like a feeding tube or surgical clips inside the esophageal lumen. Figure 1 illustrates some of the challenging cases. These difficulties lead to a high degree of uncertainty associated with the target volume of the tumor, especially at the cranial and caudal border of the tumor [4]. In order to overcome these complexities, physicians integrate CT imaging with the clinical history, endoscopic findings, endoscopic ultrasound, and other imaging modalities such as positron-emission tomography (PET) [5]. Obtaining these additional modalities is however a time-consuming and expensive process. Moreover, the process of manual delineation is a repetitive and tedious task, and often there is a lack of consensus on how to best segment the tumor from normal tissue. Despite using additional modalities and expert knowledge, the process of manual delineation still remains an ill-posed problem [6]. Nowee *et al.* [4] assessed manual delineation variability of gross tumor volume (GTV) between using CT and combined F-fluorodeoxyglucose PET (FDG-PET) [7] and CT in esophageal cancer patients in a multi-institutional study by 20 observers. They concluded that the use of PET images can significantly influence the delineated volume in some cases, however its impact on observer variation is limited.

In this paper we aim to investigate if a simplified clinical workflow based on CT scans alone allows to automatically delineate the esophageal GTV with acceptable quality. Recently, there has been a revived interest in automating this process for both the esophagus and the esophageal tumor based on CT images alone [8], [9], [10]. Our earlier work [11], leveraged the idea of dense blocks proposed by [12], arranging them in a typical U-shape. In that study, the proposed method was trained and tested on 553 chest CT scans from 49 distinct patients and achieved a DSC value of 0.73 ± 0.20, and a 95% mean surface distance (MSD) of 3.07 ± 1.86 mm for the test scans. Eight of the 85 scans in the test set had a DSC value lower than 0.50, caused by the presence of air cavities and foreign bodies in the GTV, which was rarely seen in the training data. In order to enhance the robustness of the network, in the present study we extended that work. The main contributions of this paper are as follows:

1) We propose an end-to-end CNN for esophageal GTV segmentation on CT scans. Different from much of the previous work, which addressed segmentation of the esophagus itself, we focus on the more challenging tumor area (GTV). The proposed method is end-to-end, without intricate pre- or post-processing steps, and uses no information in addition to the CT scans;

2) We introduce dilated dense attention blocks which leverage spatial and channel attention to emphasize on the GTV related features. Also, dilation layers are used to support an exponential expansion of the receptive field while keeping the size of the network fixed;

3) We collected a dataset of 228 distinct patients

(792 scans). The dataset includes different varieties of anatomies, and presence of foreign bodies and air pockets in the esophageal lumen. In this study, all patients received either Neoadjuvant or definitive chemoradiotherapy treatment options. To the best of our knowledge, none of the related works have addressed such a comprehensive dataset.

The initial results of this work were presented in [11]. The current paper includes a larger and more diverse dataset, and more elaborate evaluation. Also, we leverage dilated convolutional layers in order to increase the receptive field without increasing the size of the network, and attention gates [13] to filter tumor relevant features.

II. RELATED WORK

Most automatic esophagus segmentation approaches have used either a shape or appearance model to guide the segmentation, where training such a guidance model is complicated. Rousson *et al.* proposed a two-stage probabilistic shortest path approach to segment the esophagus from 2D CT images [6]. In the first stage, the aorta and left atrium are segmented and then registered to reference shapes in order to find a region of interest (ROI). In the second stage, the optimal esophagus centerline is extracted using the shortest path algorithm. Fieselmann *et al.* proposed an automatic approach for segmenting the esophagus by detecting the air cavities that often constitute the esophagus [14]. For reducing the time complexity, they confined all the computations to an ROI. Also, they proposed another method based on spatially-constrained shape interpolation in order to segment the esophagus in 2D CT images [15]. In that investigation, two assumptions are considered: the shape of the esophagus changes smoothly, and there is no intersection between the esophagus and the other organs.

In [16] a multi-step approach based on probabilistic models has been proposed to segment the esophagus on 3D CT scans. In that work, a pre-processing step is used to extract an ROI. Then, a discriminative learning technique is applied to label the voxels. In [17] an optical flow approach for semi-automated segmentation of CT images is used, where manually drawn curvature points are extended to contours by Fourier interpolation and afterwards, optical flow is used for registering the original contour to the other slices. This method is not only highly user-interactive but it also fails when the region to contour is topologically different between two slices. Feulner *et al.* proposed a multi-step approach based on probabilistic models for automatic segmentation of the esophagus in 3D CT scans [18]. In that work, by running a discriminative model for each axial slice, a set of approximated esophagus contours is extracted. Then, the contours are clustered and merged and afterwards, a Markov chain model is used for finding the most probable path through the axial slices. Ultimately, another discriminative model is used for refining the result. This approach just works for a manually selected ROI. The
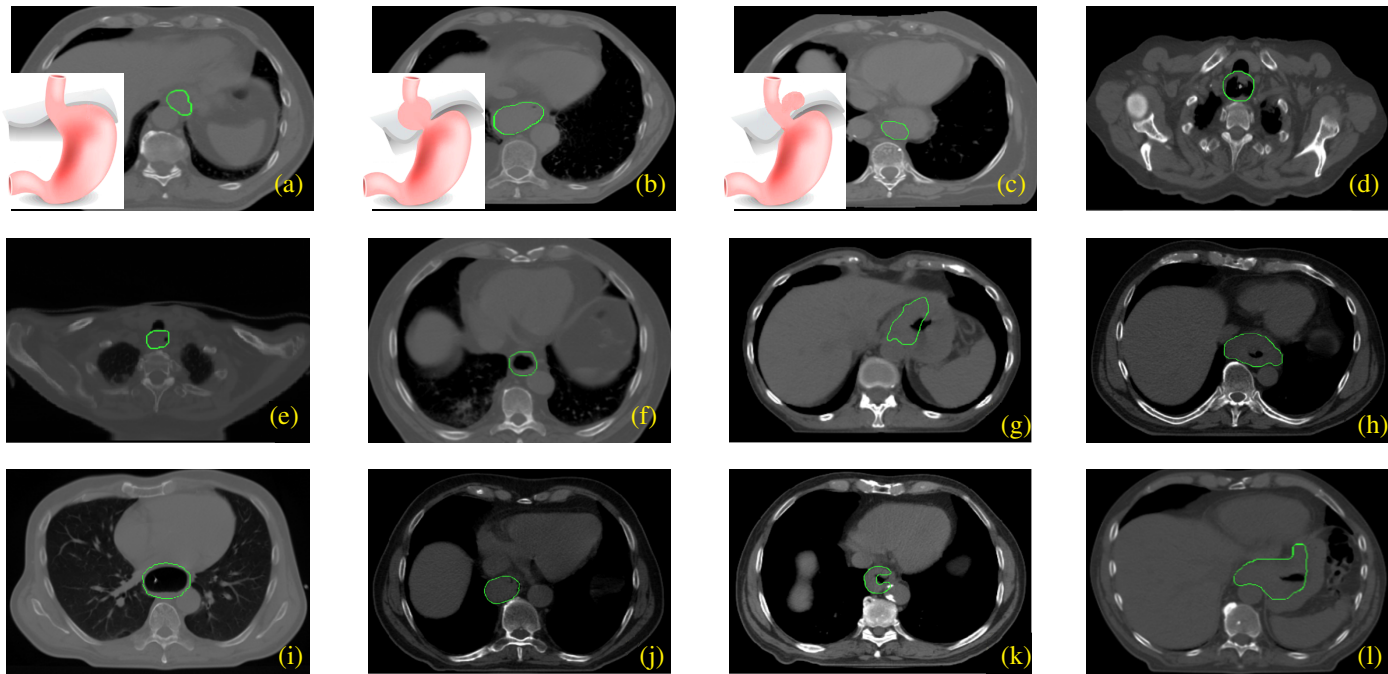
FIGURE 1: Variations in shape and location of the tumor. Green contours show the manual delineation of the GTVs. (a) normal junction of esophagus and stomach, (b) hiatal hernia (type I): migration of esophagogastric junction through the gap in the cranial direction, (c) hiatal hernia (type II): migration of esophagogastric junction in the chest, (d) proximal tumor including an air pocket and feeding tube, (e) proximal tumor including an air pocket, (f) tumor including an air pocket, (g,h) junction tumor (extension of the tumor into the stomach) including an air pocket, (i) tumor including an air pocket and feeding tube, (j) relocation of the esophagus to the left of aorta, (k) a variety in the shape of the tumor, (l) junction tumor including an air pocket.

manual selection of an ROI was later extended to automatic ROI detection by a salient landmark on the chest [19].

Kurugol *et al.* presented a 3D level set model for segmenting the esophagus over the entire thoracic range employing a shape model, with a global and a locally deformable component [20]. In their work, an initial centerline estimation is required where an ad-hoc centerline estimator was used, which was only performed at the ROI of some predefined anatomical landmarks followed by interpolation for the remaining slices. Later, they extended their work by using prior spatial and appearance models estimated from the training set instead of using the ad hoc estimator [21]. In [22], a two-phase online atlas-based approach was proposed to rank and select a subset of optimal atlas candidates for segmentation of the esophagus on CT scans. Atlas-based approaches face some restrictions including the selection of optimal atlases and a correct representation of the study population.

Deep learning for medical image analysis has aroused broad attention in recent years [23], [24], [25]. However, this technique has been limitedly used for esophagus segmentation and even less for esophageal tumor segmentation. In [26], a fully convolutional neural network (FCNN) for segmenting the esophagus on 3D CT was proposed, sur-

rounded the bottom-most of the heart and the topmost of the stomach. For refining the results, an active contour model and a random walker were used as post-processing steps. In that study, 50 scans were used as the training set and 20 scans as the test set. An average DSC value of 0.76 ± 0.11 for the test set was reported. A semi-automatic two-stage FCNN for 2D esophagus segmentation has been proposed by Trullo *et al.* [27]. The first stage performs a multi-organ segmentation in order to extract an ROI including the esophagus. Then the manually cropped ROI is fed to the second network to segment the esophagus. A DSC value of 0.72 ± 0.07 has been reported for this network with 25 scans as the training set and 30 scans as the testing set. For extracting the largest possible tumor region in 2D CT scans an FCNN was used by Hao *et al.* [23]. Then they applied a graph cut for segmenting the tumor. They reported an average DSC value of 0.75 ± 0.04 for the four patients in the test set. Jin *et al.* [8] introduced a spatial-context encoded deep esophageal clinical target volume (CTV) delineation framework to produce superior margin-based CTV boundaries. That work in an expensive pre-processing step encodes spatial context by computing the signed distance transform maps (SDMs) of the GTV, lymph nodes (LNs) and organs at risks (OARs) and then

feeds the results with the CT image into a 3D CNN. In another work Jin *et al.* [28] proposed a two-stream chained 3D CNN fusion pipeline to segment esophageal GTVs using both CT and PET+CT scans. They evaluated their approach by conducting a 5-fold cross validation on scans of 110 patients. They reported that using PET images as complementary information can improve the DSC score from 0.73 ± 0.16 to 0.76 ± 0.13. Although reasonable results can be obtained in the approaches mentioned earlier, the problem of esophageal GTV segmentation in CT modalities without extra knowledge constraining the problem is known as an ill-posed problem and remains challenging [9].

Most of the mentioned works addressed esophagus segmentation and not esophageal tumor segmentation. However, esophageal tumor segmentation is a more complicated task due to the poor contrast of the tumor with respect to its adjacent tissues. It is especially difficult to define the start and end of the tumor without additional information such as endoscopic findings.

III. THE PROPOSED METHOD

A. NETWORK ARCHITECTURE

Figure 2 shows a schematic of the proposed network, dubbed dilated dense attention Unet (DDAUnet). The network is composed of three levels, a down-sampling path for extracting contextual features and an up-sampling path for retrieving the lost resolution during extraction. In each level, different from our prior work [11], we used dilated dense spatial and channel attention blocks (DDSCAB) which is composed of a dilated dense block (DDB) and a spatial attention gate and a channel attention gate which are denoted SpA and ChA1. Using loop connectivity patterns between the layers in the DDSCAB blocks provide a deep supervision by re-using the feature maps, while dilated layers increase the receptive field exponentially without any additional parameters. Spatial attention gates are used in the main building blocks, and encourage the network to concentrate on extracting features from the tumor adjacency. The channel attention gates are used in the skip connections between the contracting and expanding paths of the Unet (named ChA2), for filtering irrelevant feature maps to improve the training process. The proposed network DDAUnet does not include ChA1, and this block is only used in some of the experiments during the optimization of the network configuration. In section IV the performance of DDAUnet will be compared with DUnet [11], dilated dense unet (DDUnet) which is DUnet with dilated convolutional layers in the dense blocks, DDAUnet without ChA2, i.e. DDAUnet-noChA2, DDAUnet with ChA1 and without SpA and ChA2, i.e. DDAUnet-noSpA-plusChA1-noChA2, DDAUnet with ChA1 and without ChA2, i.e. DDAUnet-plusChA1-noChA2.

According to [29], the incorporation of a stack of convolutional layers with small receptive fields in the first layers rather than few layers with large receptive fields decreases the number of the parameters, increases non-

linearity of the network, and consequently makes training of the network easier. These layers aid the network to extract significant features before applying convolutional operations with a wider receptive field in DDSCAB. Therefore, the network starts with two consecutive $(3 \times 3 \times 3)\mathrm{conv}_{d=1,p=true} + \mathrm{BN} + \mathrm{ReLU}$, in which $3 \times 3 \times 3$, d and p indicate the kernel size, dilation factor and padding of the convolutional layer respectively. Also, BN and ReLU denote batch normalization and a Rectified linear unit layer, respectively.

Afterward, the network is followed by a DDSCAB composed of a dilated dense block (DDB) and spatial and channel attention gates. For each DDB, R is the number of sub-DDBs. In each sub-DDB, there are R number of $(3 \times 3 \times 3)\mathrm{conv}_{d=2,p=true} + \mathrm{BN} + \mathrm{ReLU}$ and R number of $(1 \times 1 \times 1)\mathrm{conv}_{d=1,p=true} + \mathrm{BN} + \mathrm{ReLU}$ layers. The output of a DDB is the concatenation of all preceding sub-DDBs. In our prior work, it has been shown that the loop connectivity patterns in dense blocks assist the network to perform better [11]. In the DDBs, $(1 \times 1 \times 1)\mathrm{conv}$ layers are used as bottleneck layers, which compress the number of feature maps and thus improve computational efficiency [12]. In this paper, the feature maps in each DDB are compressed by a compression coefficient of θ. The output of DDB then is fed to spatial and channel attention gates in order to selectively filter the GTV irrelevant spatial features and feature maps respectively, which leads to improving the training process. In the down-sampling path, the DDSCABs are followed by $(1 \times 1 \times 1)\mathrm{conv}_{d=1,p=true} + \mathrm{BN} + \mathrm{ReLU}$. Using $1 \times 1 \times 1$ convolutional layers does not affect the receptive field of the network, however, increases the non-linearity in between layers [30]. At the end of down-sampling path and in the up-sampling path every DDSCAB is followed by $(3 \times 3 \times 3)\mathrm{conv}_{d=1} + \mathrm{BN} + \mathrm{ReLU}$. In Section IV we will investigate the effect of deploying spatial and channel gates in DDSCAB and will see that utilizing only the spatial gate is more effective. Also, the skip connections between the contracting and expanding path are equipped by channel attention gates to filter the irrelevant feature maps. Later we will show that leveraging the spatial and channel attention gates aid the network to end up with better segmentation results. The network is finalized by a convolutional layer with linear activation and a soft-max layer to compute a probabilistic output. The probabilistic output can be classified as tumor and non-tumor regions. The skip connections between the down-sampling and up-sampling paths demonstrate cropped concatenation of the feature maps of the corresponding down-sampling levels and up-sampling levels.

B. LOSS FUNCTION

In this work, similar to our prior work [11] we used the Dice coefficient as our main loss function [31]:

$$\mathrm{DSC}_{\mathrm{GTV}} = \frac{2\sum_i^N s_i g_i}{\sum_i^N s_i^2 + \sum_i^N g_i^2}, \qquad (1)$$
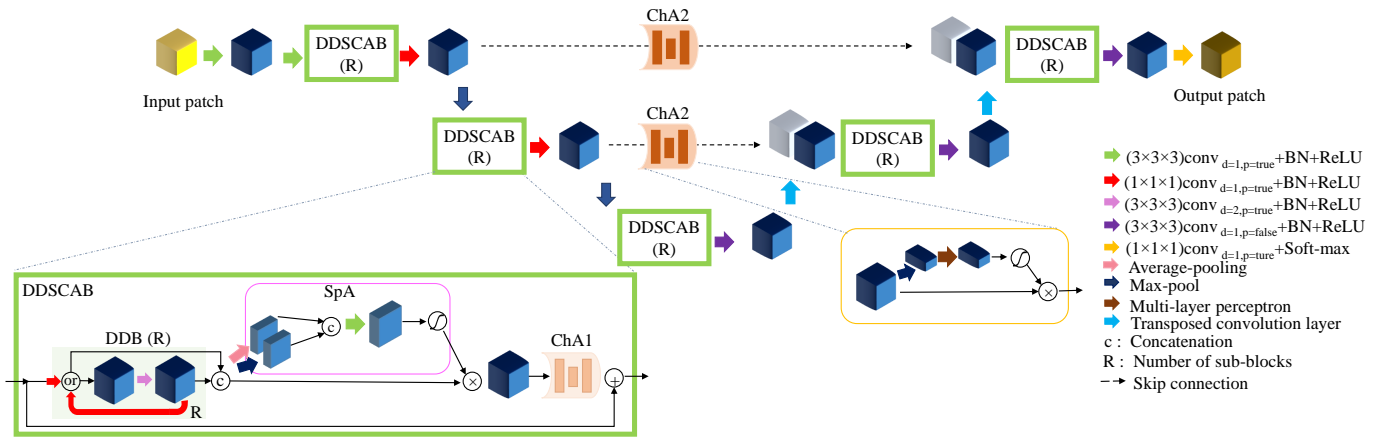
FIGURE 2: The architecture of the proposed method. DDSCAB and DDB stand for dilated dense spatial and channel attention block and dilated dense block, respectively. R is the number of sub-DDBs. ChA1, ChA2, and SpA denote channel attention gate located on skip connections, channel attention inside the DDSCAB block, and spatial attention gates inside DDSCAB. ChA1, shown transparently here, is not included in the final network (DDAUnet), but is used in some of the experiments.

where $s_i \in S$ is the binary segmentation of the GTV predicted by the network and $g_i \in G$ is the ground truth segmentation. We investigated different combinations of loss functions including boundary loss [32], distance map loss [33], and focal Dice [34]. In [32] it is shown that the boundary loss can be approximated by:

$$\mathcal{L}_B(\theta) = \int_\omega \phi_G(q) s_\theta(q) dq, \qquad (2)$$

where ϕ_G and s_θ denote the level set representation of the boundary of ground truth, and network output, respectively. The traditional loss functions like Dice are based on integrals over the segmentation regions [32]. In this work we leverage the boundary loss function to aid the network to also focus on the outer contours. In Section IV it will be discussed that the combination of Dice and boundary loss works the best for this problem.

IV. DATA, TRAINING DETAILS AND EVALUATION
A. DATASET
All patients of this study received one of the following two treatments:

(A) Neoadjuvant chemo-radiotherapy (CRT) followed by surgical resection. The radiotherapy is 23×1.8 Gy, 5 fractions a week. The external beam radiotherapy consisted of 23 fractions of 1.8 Gy, five fractions per week. Concurrent chemotherapy consisted of 5 weekly administrations of carboplatin and paclitaxel.

(B) If patients are inoperable (proximal tumors they receive primary/definitive chemoradiotherapy (CRT). The external beam radiotherapy consisted of 28 fractions of 1.8 Gy, five fractions per week. Concurrent chemotherapy consisted of 6 weekly administrations of carboplatin and paclitaxel.

The dataset used in this study consists of 288 distinct patients acquired for a study approved by the Medical Ethics Review Committee of Leiden University Medical Center, the Netherlands. The dataset includes two sub-datasets from 21 and 267 patients, respectively, in which each patient received either treatment plan A or B. The data acquisition was performed with a Brilliance Big Bore scanner (Philips Healthcare, Best,Holland) and the delineation process was done by Pinnacle3, (version 9.6–9.8; Philips Radiation Oncology Systems, Fitchburg, WI.) treatment planning software. The ground truth segmentation was performed by the MD on 2D axial slices and evaluated on the 3D cardinal planes. Table 1 tabulates the details of the datasets. The first dataset includes five repeat CT scans acquired at different time points. Three time-points contain only one 3D CT scan, and two time-points include one 3D CT scan and one 4D CT scan, composed of 10 breathing phases. Each subdataset includes a corresponding esophageal GTV segmentation for each CT scan, which has been delineated by one (dataset I) or multiple (dataset II) experienced physicians. Each scan contains 58-108 slices with an image resolution of 512×512 pixels and an average voxel thickness of $0.98 \times 0.98 \times 3 \text{mm}^3$, and were re-sampled to a voxel size of $1 \times 1 \times 3 \text{mm}^3$ in this study.

B. TRAINING DETAILS
In this work, the proposed network, which contains 651,098 trainable parameters, has been implemented in Google's Tensorflow and the experiments are carried out using a NVIDIA Quadro RTX6000 with 24 GB of GPU memory. For all networks, the patch extraction process has been implemented by multi-threaded programming in which fetching the images into RAM, extracting the patches from the fetched images and feeding the extracted patches to the GPU are done concurrently. The multi-threading technique

TABLE 1: Details of the dataset

dataset	# of patients	# of scans	Type	Time period	Treatment plan
I	21 (A)	525	5 time-points 2 time-points: 3D scans 3 time-points: 3D + 4D scans	2012-2014	A: Neoadj. CRT
II	162 (A) + 105 (B)	267	3D	2014-2019	A: Neoadj. CRT B: def. CRT

speeds up the patch extraction process. The input patches have been augmented by white noise extracted from a Gaussian distribution of $N(\mu', \sigma')$, in which $\mu' = 0$ is the mean of the distribution and σ' is the standard deviation, which is selected randomly between 0 and 5. During the test process, the fully convolutional nature of the network is used, with zero padding to yield equal output size. For managing the GPU memory with a larger input patch, we use a batch size of seven.

For designing the best configuration of the network, we perform experiments comparing different architectures and loss functions. Since the size of our dataset was relatively large with 792 scans, this led to a training time of 6 days. For this reason, similar to [35], [24], [36], we adopted a 3-fold Monte Carlo cross-validation approach [37] in this study. Hence, the datasets I and II are divided randomly into three distinct sets detailed in Table 2. The optimization of the network is performed on the validation set. The test set is excluded from the model optimization and kept independently for the final evaluation. After choosing the best configuration of the network, the final model is trained for two more random splits of the training and validation sets, resulting in three trained models. At the end, an average of the final results for the chosen network, trained on three splits, is reported on the test set.

In Section V the optimization of the network configuration will be discussed on the validation set. Then the best network is trained by different linear combinations of the loss functions including Dice, boundary loss, distance map, and Focal loss. Then in Section V-B, for reproducibility, the results of the best configuration of the network will be reported for two more distinct and random splits.

TABLE 2: Data split into training, validation and testing sets. P and S denote distinct patients and scans.

dataset	# of P/S	DB I	DB II	total
training	P	13	182	195
	S	325	182	507
validation	P	2	23	25
	S	50	23	53
testing	P	6	62	68
	S	150	62	212
total	P	21	267	288
	S	525	267	792

C. EVALUATION MEASURES

For evaluating the results we report DSC value (see Section III-B), MSD and Hausdorff distance (HD) which are

defined as:

$$\text{MSD} = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} d(a_i, S) + \frac{1}{m}\sum_{i=1}^{m} d(b_i, S)\right), \quad (3)$$

$$\text{HD} = \max\{\max_i\{d(a_i, S)\}, \max_j\{d(b_i, G)\}\}, \quad (4)$$

in which S and G are the predicted and ground truth contours, and $\{a_1, ..., a_n\}$ and $\{b_1, ..., b_m\}$ the surface mesh points of S and G, and $d(a_i, S) = min_j\|b_j - a_i\|$ respectively. For the Hausdorff distance, we report the 95% percentile instead of the maximum for robustness against outliers. Since defining the slices where the tumor starts and stops is difficult even for medical doctors, we report perpendicular cranial and caudal distance between the output of the CNNs and the ground truth. The cranial distance (CrD) error is computed as the topmost slice number of the ground truth minus the topmost slice number of the CNN prediction; the caudal distance (CaD) error is computed similarly.

V. EXPERIMENTAL RESULTS

In this section the experimental results are reported, with the datasets divided into training, validation and test sets as described in Section IV-B. Model optimization experiments are described in Section V-A, where comparison is performed on the validation set. Subsequently, the final results are reported on the test set in Section V-B. For all experiments, we extract the largest component of the network prediction using connected component analysis, and report that. A repeated measure oneway ANOVA test was performed on the Dice values using a significance level of p = 0.05.

A. MODEL OPTIMIZATION

We explored the effect of combinations of R values on the results, where R is the number of sub-dilated dense blocks in the network. Figure 3 shows the boxplots of DSC, MSD, 95%HD, cumulative frequency (%) of DSC, and perpendicular cranial and caudal distance errors for different combinations of R. Since the setup $R = [3, 3, 5, 3, 3]$ performed best, this configuration was selected for the remainder of the paper.

Figure 4 shows the results for different configurations of the CNN models using the DSC loss function. Since channel attention gates inside the DDSCAB block, i.e. ChA1 in Fig. 2, did not improve the results, these are not used in the final configuration. The results show that DDAUnet outperforms the other network configurations significantly.

Figure 5 shows the loss curves for different network architectures on the validation set during the training process,
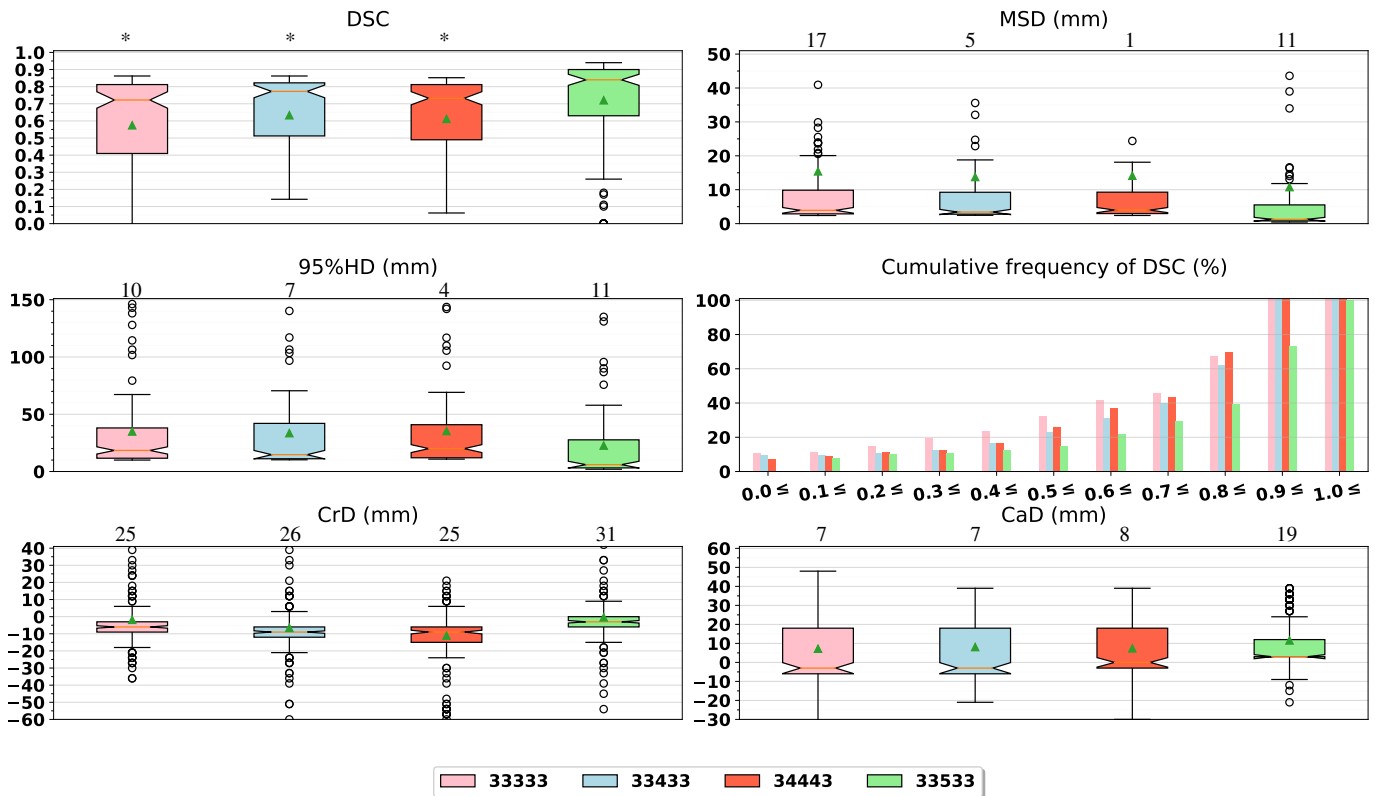
FIGURE 3: A comparison between different dense block configurations for DDAUnet on the validation set. Each number in the configuration setting shows the number of conv-BN-ReLU in the dense blocks of the network. The number of results with values larger than the maximum value on the vertical axis, is shown on top of each plot. The stars in the DSC plot show statistical significance between 33533 and the other configurations.

TABLE 3: AUC for the networks on the validation set

model	AUC
DUnet	0.49
DDUnet	0.53
DDAUnet-noChA2	0.73
DDAUnet-plusChA1-noChA2	0.63
DDAUnet-noSpA-plusChA1-noChA2	0.71
DDAUnet	0.76

smoothed with a kernel of size 3. It can be seen that the networks equipped with attention gates converged quicker. Figure 6 shows the precision-recall curves for the networks on the validation set using the DSC loss function. The precision and recall were calculated with different threshold values applied to the probabilistic output of the networks. For acquiring the final segmentation we used a threshold of 0.5. Table 3 tabulates the values of the area under the curve (AUC) for the networks on the validation set [38]. The AUC for DDAUnet is the largest, and we choose this method as the final network architecture.

We experimented with different combinations of loss functions including Dice, boundary loss, distance map loss, Focal Dice on the validation set. Figure 7 shows the results. The results show that DDAUnet using the DSC + BL loss function outperforms the other loss functions significantly.

B. FINAL RESULTS

As explained before, repeatability and reproducibility studies were conducted for three distinct and random splits of training and validation sets. Table 4 shows the results on the independent test set after applying the largest component analysis. Figure 8 shows example results of the final network for some patients with different shape varieties and difficulties raised by the presence of air pockets or feeding tubes. The 2D DSC values are shown in yellow. Figure 9 shows a qualitative comparison between the different CNNs.

In order to study the strengths and weaknesses of the final model, we manually labeled each scan with the following properties: presence of air pockets in the esophagus, the presence of a feeding tube in the esophageal lumen, the tumor is a junction tumor, the tumor volume is larger than 30cc (which is defined by the median split of the GTV volumes), the patient has a hiatal hernia, the tumor is in a dislocated esophagus, the tumor is located in the proximal esophagus (proximal tumor). Figure 10 shows the results of DSC value, MSD and 95%HD for the mentioned tags for the final network on the test set. Results show that the network works better for patients with absence of air pockets, feeding tubes, or junction tumors. This may be caused by the different varieties raised by the existence of
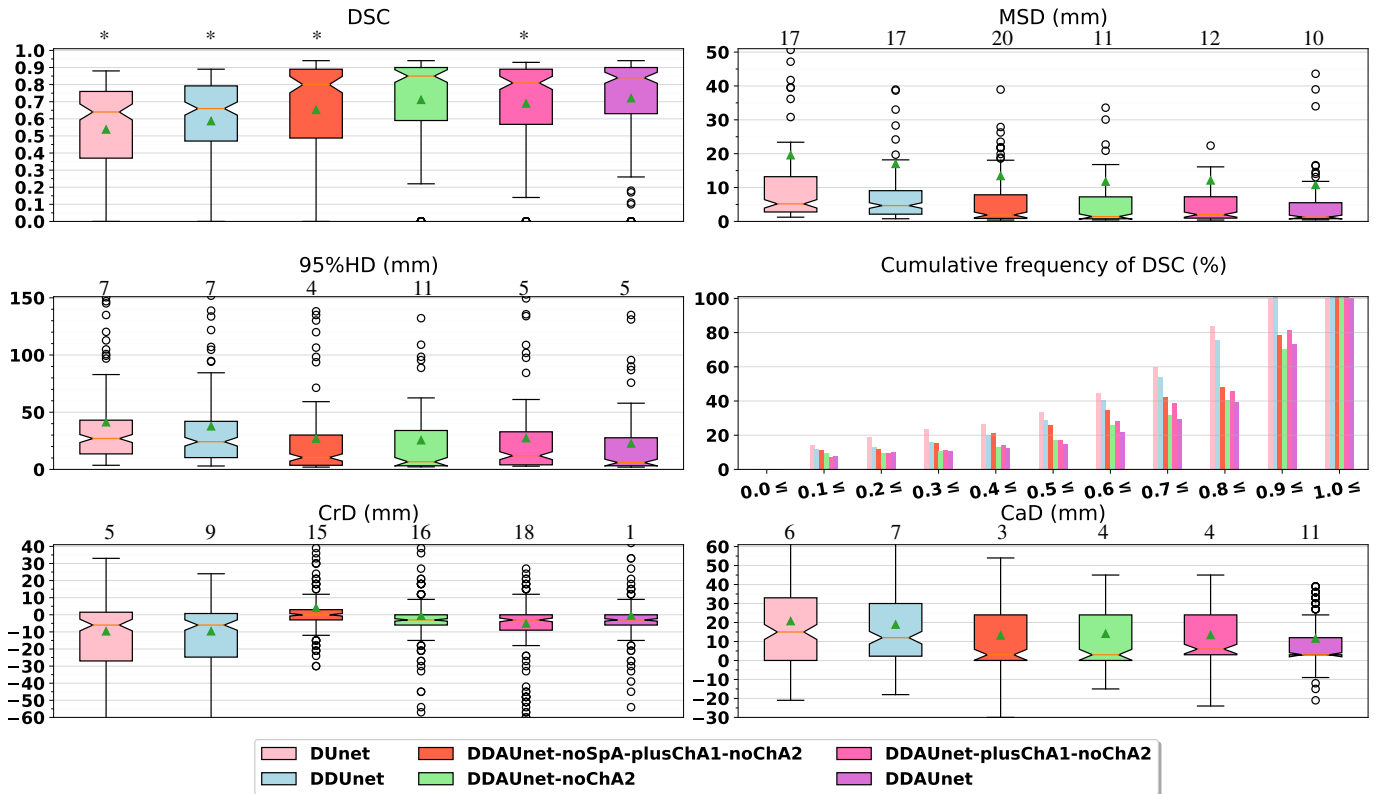
FIGURE 4: A comparison between different network configurations on the validation set. The number of results with values larger than the maximum value on the vertical axis, is shown on top of each plot. The stars in the DSC plot show statistical significance between DDAUnet and the other CNNs.

TABLE 4: Results for DDAUnet on the independent test set, with the combined Dice and boundary loss function.

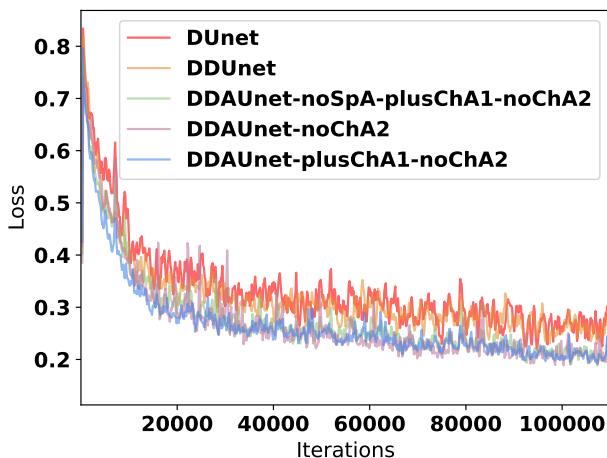| Split	DSC $\mu \pm \sigma$	CrD (mm) $\mu \pm \sigma$	CaD (mm) $\mu \pm \sigma$	MSD (mm) $\mu \pm \sigma$	95%HD (mm) $\mu \pm \sigma$
1	0.79 ± 0.20	-6.4 ± 26.0	3.1 ± 12.8	6.2 ± 23.2	16.1 ± 28.1
2	0.79 ± 0.19	-8.6 ± 17.7	4.6 ± 13.1	4.6 ± 16.2	14.6 ± 22.2
3	0.78 ± 0.21	-4.6 ± 12.7	2.9 ± 10.8	5.5 ± 20.6	13.5 ± 24.3
Mean	0.79 ± 0.20	-6.5 ± 19.6	3.5 ± 12.3	5.4 ± 20.2	14.7 ± 25.0

FIGURE 5: Loss curves for the validation set during the training process.

air pockets or foreign bodies. Also, the number of patients with a hiatal hernia, a dislocated esophagus or a proximal tumor is relatively small in the test set, not allowing to draw a conclusion.

VI. DISCUSSION

Esophageal GTV segmentation is not a trivial problem, due to the difficulties raised by the poor contrast with respect to its vicinity. Most research addressed only segmentation of the esophagus, while esophageal GTV segmentation has been touched in few works. Since defining a correct start and end location (slice) of the tumor in the cranial-caudal direction based on CT images alone, is not an easy task even for doctors, esophageal GTV segmentation is considered as an ill-posed problem. In this paper, for addressing the esophageal GTV segmentation we designed an efficient deep learning model.

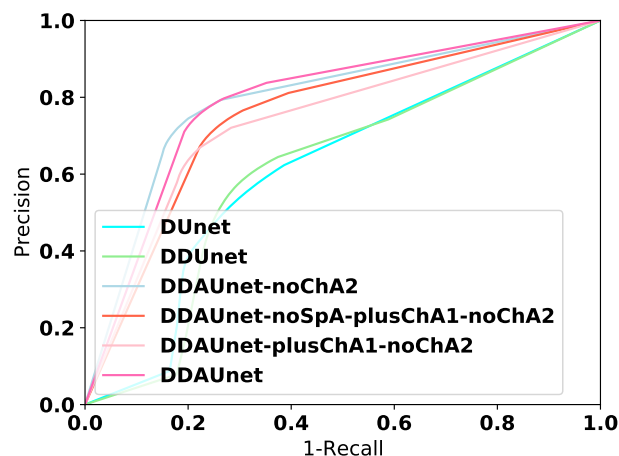In terms of the training data, we collected 792 CT scans

FIGURE 6: Precision-recall curves for the different network architectures on the validation set using the DSC loss function.

from 288 patients diagnosed with an esophageal tumor. This dataset is the largest dataset among the present works addressing esophageal tumor segmentation. Training time for a single network was in the order of 5 days. The average inference time for the final network for a cube of $255 \times 255 \times 255$ voxels, is 4.0 ± 1.1 seconds.

For tuning the proposed network, many experiments were performed in the present paper. We leveraged the DenseUnet network, already deployed in our prior work as a baseline. In order to increase the receptive field of the network, dense blocks were equipped by dilated convolutional layers, dubbed the DDUnet network. We leveraged attention mechanisms to encourage the network to selectively filter out GTV irrelevant features. Three types of attention gates were utilized: i) a spatial attention gate in the dense blocks to filter out GTV irrelevant features in the spatial domain of each feature map, ii) a channel attention gate in the dense blocks (ChA1) to filter out irrelevant feature maps entirely, and iii) skip attention gates (ChA2) to filter out GTV irrelevant feature maps between the contracting and expanding paths of the Unet. The experiments on the validation set showed that the architecture with the spatial attention and skip attention gates (ChA2), dubbed DDAUnet, achieved the best result. The use of attention gates in the dense blocks (ChA1) did not improve the results, compared to the baseline model. This may be caused by the removal of feature maps in early levels of the network, subsequently preventing the network to extract fine features at deeper levels. On the other hand, the ChA2 gates filter out redundant or irrelevant feature maps during the retrieval of lost resolution. This leads the network to prevent to concatenate tumor-irrelevant feature maps and consequently aids the network to utilize its capacity on the tumor region. The optimized network architecture was further tuned using a large variety of loss functions, again on the validation set. Results showed that the summation of Dice and boundary loss performed

best. Combining the two loss functions allows the network to concentrate not only on the region but also on the outer contours, thereby improving the results. This confirms results from [32] where this improvement was also observed. Therefore, we introduced the DDAUnet with the summation of Dice and boundary loss as the loss function in the final network.

We trained the final network for three random splits of the training and validation sets. The results on the test set showed an average DSC of 0.79 ± 0.20, an MSD of $5.4 \pm 20.2mm$, a 95%HD of $14.7 \pm 25.0mm$, and cranial and caudal perpendicular distance errors of $-6.5 \pm 12.3mm$ and $5.4 \pm 20.2mm$, respectively. The cranial and caudal perpendicular distance errors between the ground truth and the network result show that the network overestimates at the top of the GTV by ~6.5 mm, and underestimates at the bottom of the GTV by ~3.5 mm. As slice thickness of the data was 3 mm, this translates to approximately 2 and 1 slices on average, respectively. For alleviating this issue, incorporating auxiliary information could aid the network.

Although the datasets are not comparable, in [28] an average DSC score of 0.76 ± 0.13 was obtained on scans of 110 patients, using 5-fold cross validation. In [23] a DSC score of 0.75 ± 0.04 for four patients as the test set has been reported. In our prior work [11], we achieved a DSC value of 0.73 ± 0.20, and a 95% mean surface distance MSD of 3.07 ± 1.86 mm for 85 CT scans from 13 distinct patients. In the work described in this paper, a higher DSC value was obtained.

Nowee *et al.* studied the inter-observer variability in esophageal tumour delineation, and found that this variability is mainly located at the cranial and caudal border [4]. They report a generalized conformity index for the GTV, a measure related to Dice overlap but for multiple observers, of 0.67. The human delineation variation in the cranial direction, defined as the standard deviation of the most proximal slice, was on average 9.9 mm, and 7.5 mm for the caudal direction. Although these measures are not the same as the measures reported in this paper, we cautiously conclude that the cranial and caudal error of the proposed automatic method (see Table 4) is not far from human delineation variation.

Nowee et al. also investigated the impact of incorporating FDG-PET scans in the delineation process, and concluded that although it can influence the delineated volume significantly, its impact on observer variation was limited. As a future work, we aim to study if fusion of CT with FDG-PET can aid the CNNs to improve the extracted features and subsequently the segmentation results.

For a close inspection, we investigated the results on the independent test set for the final network. We labelled the patients in the test set with different tags, including the presence of air pockets, feeding tube, junction tumor, tumor volume $> 30cc$, hiatal hernia, dislocated esophagus, proximal tumor. Inspection of the final results (see Figure 10) showed that the network performed better for patients
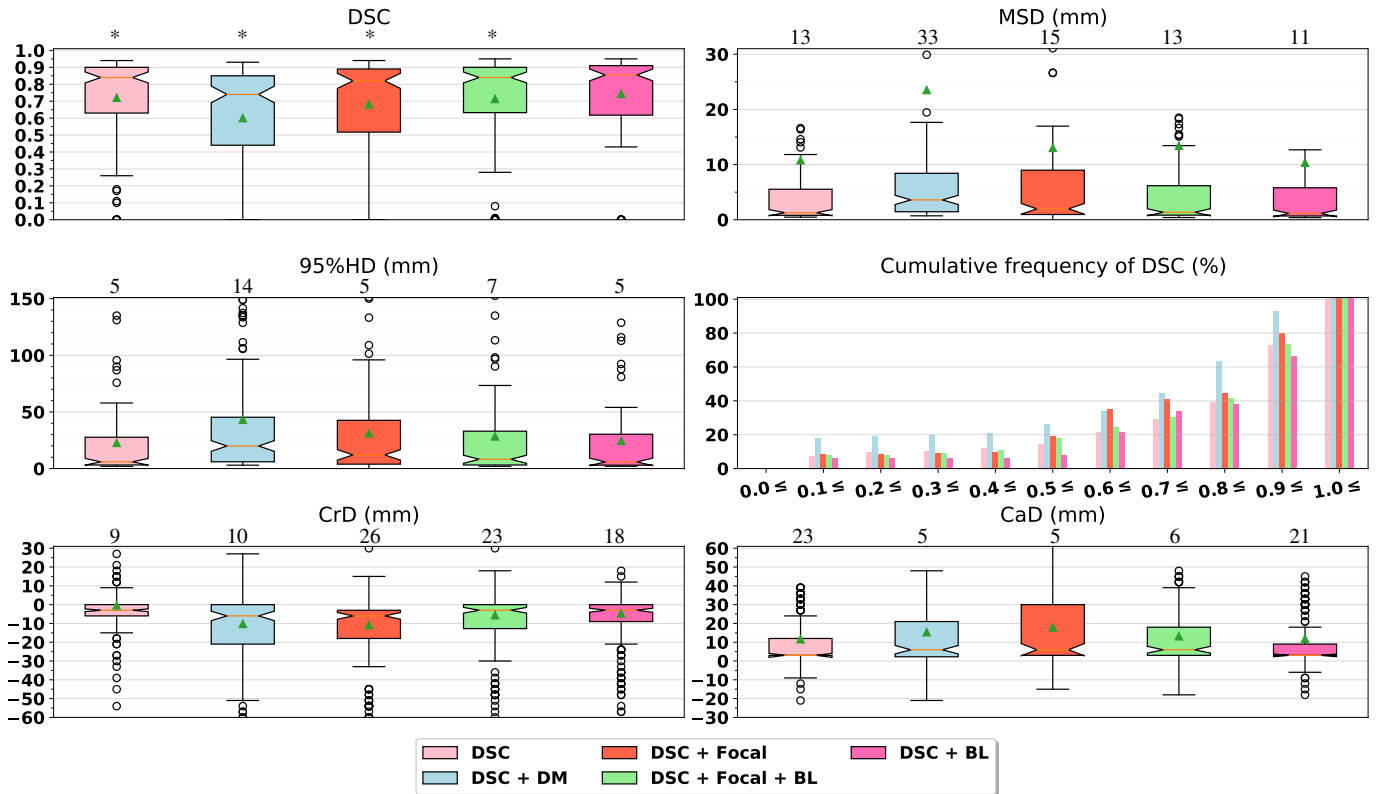
FIGURE 7: A comparison between deploying different loss functions for DDAUnet on the validation set. The number of results with values larger than the maximum value on the vertical axis, is shown on top of each plot. The stars in the DSC plot show statistical significance between DSC+BL and the other loss functions.

with an absence of air pockets, feeding tubes in the esophagus lumen, or junction tumors. A lower performance was obtained for smaller tumors ($< 30cc$), while the strength of the network for patients with a dislocated esophagus, a proximal tumor, or a hiatal hernia was not judge-able. Therefore, enriching the dataset with more patients with the mentioned properties would potentially improve the performance of the model. Also, incorporating endoscopic findings in the process of segmentation can be considered as a future work to investigate if that can aid CNNs to reduce errors specially at the start and end of the GTV.

VII. CONCLUSION

In this study, we collected a large set of CT scans from 288 distinct patients with esophageal cancer. To the best of our knowledge this is the largest dataset in esophageal tumor segmentation literature to date. We showed that despite the difficulties raised by poor contrast of esophageal tumors with respect to their neighbouring tissues, varieties in shape and location of tumor, presence of air pockets and foreign bodies, the proposed method, dubbed dilated dense attention Unet (DDAUnet), could segment the gross tumor volume with a mean surface distance of $5.4 \pm 20.2mm$.

VIII. ACKNOWLEDGMENTS

We gratefully acknowledge Femke P. Peters for collection and delineation of the data.

REFERENCES

[1] Peter C Enzinger and Robert J Mayer. Esophageal cancer. New England Journal of Medicine, 349(23):2241–2252, 2003.

[2] Jacques Ferlay, Isabelle Soerjomataram, Rajesh Dikshit, Sultan Eser, Colin Mathers, Marise Rebelo, Donald Maxwell Parkin, David Forman, and Freddie Bray. Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012. International Journal of Cancer, 136(5):E359–E386, 2015.

[3] Marcelo Mamede, Paula Abreu-e Lima, Maria Raquel Oliva, Vânia Nosé, Harvey Mamon, and Victor H Gerbaudo. FDG-PET/CT tumor segmentation-derived indices of metabolic activity to assess response to neoadjuvant therapy and progression-free survival in esophageal cancer: correlation with histopathology results. American Journal of Clinical Oncology, 30(4):377–388, 2007.

[4] Marlies E Nowee, Francine EM Voncken, Alexis NTJ Kotte, Lucas Goense, Peter SN van Rossum, Astrid LHMW van Lier, Stijn W Heijmink, Berthe MP Aleman, et al. Gross tumour delineation on computed tomography and positron emission tomography-computed tomography in oesophageal cancer: A nationwide study. Clinical and Translational Radiation Oncology, 14:33–39, 2019.

[5] Thomas R Charles, John G Hunter, and Blair AA Jobe. Esophageal cancer: principles and practice. Demos Medical Publishing, 2009.

[6] Mikael Rousson, Ying Bai, Chenyang Xu, and Frank Sauer. Probabilistic minimal path for automated esophagus segmentation. In Medical Imaging 2006: Image Processing, volume 6144, page 614449. International Society for Optics and Photonics, 2006.

[7] Frederiek M Lever, Irene M Lips, Sjoerd PM Crijns, Onne Reerink, Astrid LHMW van Lier, Marinus A Moerland, Marco van Vulpen, and

(a) Normal esophageal GTV

(b) Normal esophageal GTV

(c) GTV including an air pocket

(d) GTV in a dislocated esophagus

(e) Junction GTV

(f) Junction GTV

(g) GTV including an air pocket

(h) Proximal GTV including an air pocket and a feeding tube

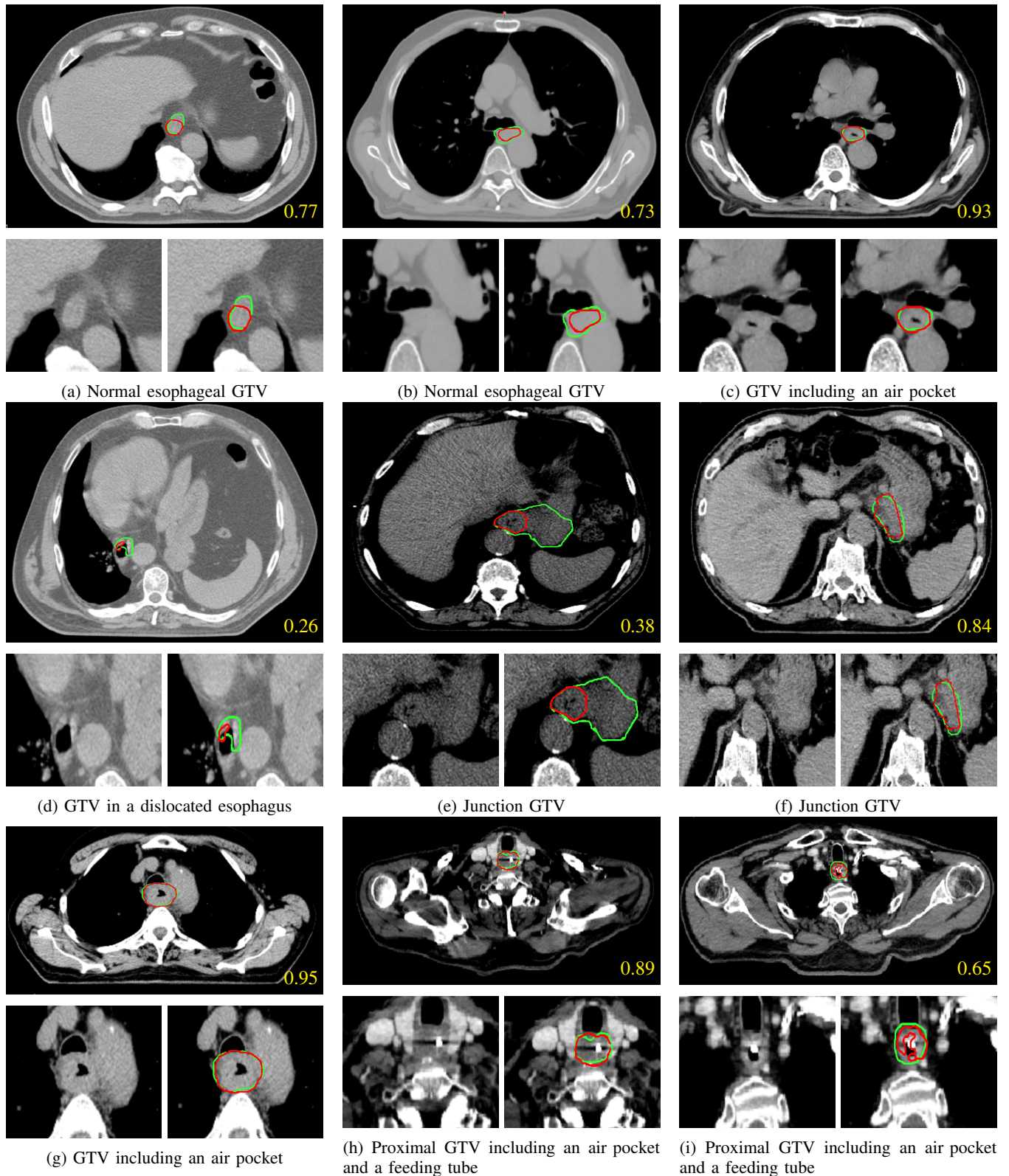(i) Proximal GTV including an air pocket and a feeding tube

FIGURE 8: Example results of the proposed method with 2D DSC value in yellow. The manual delineation and the network results are shown by green and red contours, respectively.
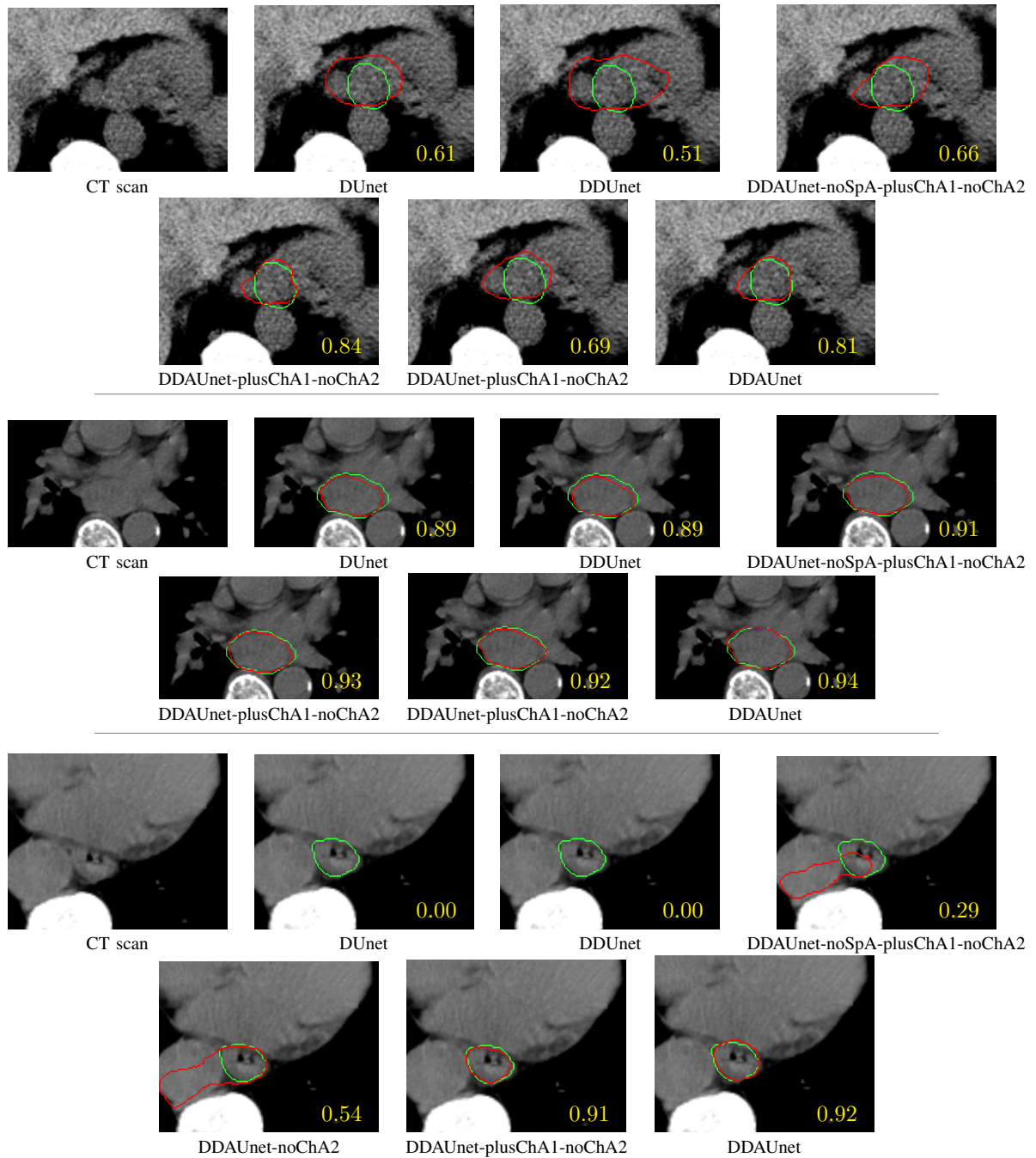
FIGURE 9: Qualitative comparison of DDAUnet with the other CNNs for three slices from three distinct patients. 2D DSC values are show in yellow. The manual delineation and the network results are shown by green and red contours, respectively.
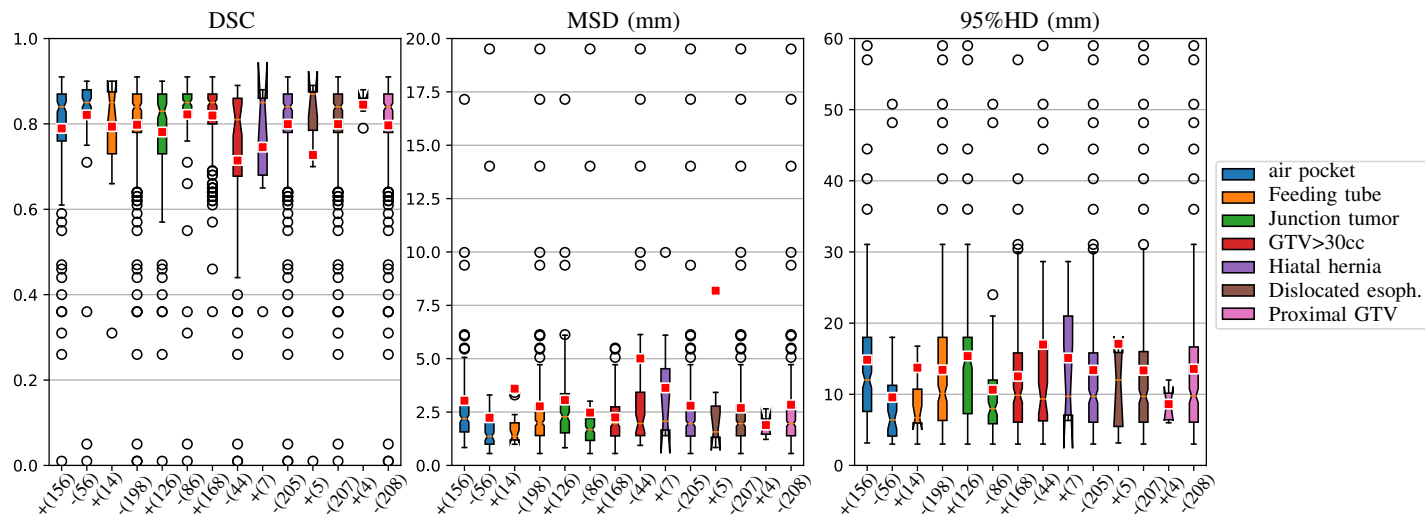
FIGURE 10: Results analysis for DDAUnet, DSC, MSD and 95%HD boxplots on the test data for different patients with or without an air pocket, a feeding tube in the esophagus lumen, a junction tumor, tumor volume larger than 30cc (which is defined by the median split of the GTV volumes), a hiatal hernia, a dislocated esophagus, or a proximal tumor. The outliers larger than 20 and 60 forMSD and 95%HD have not been shown. The number of scans for each boxplot is shown in parentheses below each plot.

Gert J Meijer. Quantification of esophageal tumor motion on cine-magnetic resonance imaging. International Journal of Radiation Oncology* Biology* Physics, 88(2):419–424, 2014.

[8] Dakai Jin, Dazhou Guo, Tsung-Ying Ho, Adam P Harrison, Jing Xiao, Chen-kan Tseng, and Le Lu. Deep esophageal clinical target volume delineation using encoded 3D spatial context of tumors, lymph nodes, and organs at risk. In MICCAI, pages 603–612. Springer, 2019.

[9] Di-Xiu Xue, Rong Zhang, Yuan-Yuan Zhao, Jian-Ming Xu, and Ya-Lei Wang. Fully convolutional networks with double-label for esophageal cancer image segmentation by self-transfer learning. In Ninth International Conference on Digital Image Processing (ICDIP 2017), volume 10420, page 104202D. International Society for Optics and Photonics, 2017.

[10] Ying Liang, Diane Schott, Ying Zhang, Zhiwu Wang, Haidy Nasief, Eric Paulson, William Hall, Paul Knechtges, Beth Erickson, and X Allen Li. Auto-segmentation of pancreatic tumor in multi-parametric MRI using deep convolutional neural networks. Radiotherapy and Oncology, 145:193–200, 2020.

[11] Sahar Yousefi, Hessam Sokooti, Mohamed S Elmahdy, Femke P Peters, Mohammad T Manzuri Shalmani, Roel T Zinkstok, and Marius Staring. Esophageal gross tumor volume segmentation using a 3D convolutional neural network. In MICCAI, pages 343–351. Springer, 2018.

[12] Gao Huang, Zhuang Liu, Kilian Q Weinberger, and Laurens van der Maaten. Densely connected convolutional networks. In CVPR, pages 4700–4708, 2017.

[13] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Proceedings of the European conference on computer vision (ECCV), pages 3–19, 2018.

[14] Andreas Fieselmann, Stefan Lautenschläger, Frank Deinzer, and Björn Poppe. Automatic detection of air holes inside the esophagus in CT images. In Bildverarbeitung für die Medizin 2008, pages 397–401. Springer, 2008.

[15] Andreas Fieselmann, Stefan Lautenschläger, Frank Deinzer, Matthias John, and Björn Poppe. Esophagus segmentation by spatially-constrained shape interpolation. In Bildverarbeitung für die Medizin 2008, pages 247–251. Springer, 2008.

[16] Johannes Feulner, S Kevin Zhou, Matthias Hammon, Sascha Seifert, Martin Huber, Dorin Comaniciu, Joachim Hornegger, and Alexander Cavallaro. A probabilistic model for automatic segmentation of the esophagus in 3-D CT scans. IEEE Transactions on Medical Imaging, 30(6):1252–1264, 2011.

[17] Tzung-Chi Huang, Geoffrey Zhang, Thomas Guerrero, George Starkschall, Kan-Ping Lin, and Ken Forster. Semi-automated CT segmentation using optic flow and fourier interpolation techniques. Computer Methods and Programs in Biomedicine, 84(2-3):124–134, 2006.

[18] Johannes Feulner, S Kevin Zhou, Alexander Cavallaro, Sascha Seifert, Joachim Hornegger, and Dorin Comaniciu. Fast automatic segmentation of the esophagus from 3D CT data using a probabilistic model. In MICCAI, pages 255–262. Springer, 2009.

[19] Johannes Feulner, S Kevin Zhou, Martin Huber, Alexander Cavallaro, Joachim Hornegger, and Dorin Comaniciu. Model-based esophagus segmentation from CT scans using a spatial probability map. In MICCAI, pages 95–102. Springer, 2010.

[20] Sila Kurugol, Necmiye Ozay, Jennifer G Dy, Gregory C Sharp, and Dana H Brooks. Locally deformable shape model to improve 3D level set based esophagus segmentation. In ICPR 2010, pages 3955–3958. IEEE, 2010.

[21] Sila Kurugol, Erhan Bas, Deniz Erdogmus, Jennifer G Dy, Gregory C Sharp, and Dana H Brooks. Centerline extraction with principal curve tracing to improve 3D level set esophagus segmentation in CT images. In EMBC 2011, pages 3403–3406. IEEE, 2011.

[22] Jinzhong Yang, Benjamin Haas, Raymond Fang, Beth M Beadle, Adam S Garden, Zhongxing Liao, Lifei Zhang, Peter Balter, et al. Atlas ranking and selection for automatic segmentation of the esophagus from CT scans. Physics in Medicine & Biology, 62(23):9140, 2017.

[23] Zhaojun Hao, Jiwei Liu, and Jianfei Liu. Esophagus tumor segmentation using fully convolutional neural network and graph cut. In Chinese Intelligent Systems Conference, pages 413–420. Springer, 2017.

[24] Shuai Chen, Gerda Bortsova, Antonio García-Uceda Juárez, Gijs van Tulder, and Marleen de Bruijne. Multi-task attention-based semi-supervised learning for medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pages 457–465. Springer, 2019.

[25] Dakai Jin, Dazhou Guo, Tsung-Ying Ho, Adam P Harrison, Jing Xiao, Chen-kan Tseng, and Le Lu. Deeptarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. Medical Image Analysis, 68:101909, 2020.

[26] Tobias Fechter, Sonja Adebahr, Dimos Baltas, Ismail Ben Ayed, Christian Desrosiers, and Jose Dolz. A 3D fully convolutional neural network and a random walker to segment the esophagus in CT. Journal of Medical Physics, 2017.

[27] Roger Trullo, Caroline Petitjean, Dong Nie, Dinggang Shen, and Su Ruan. Fully automated esophagus segmentation with a hierarchical deep learning approach. In ICSIPA 2017, pages 503–506. IEEE, 2017.

[28] Dakai Jin, Dazhou Guo, Tsung-Ying Ho, Adam P Harrison, Jing Xiao, Chen-kan Tseng, and Le Lu. Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion. In MICCAI, pages 182–191. Springer, 2019.

[29] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene under-standing. In CVPR, pages 3213–3223, 2016.

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional net-works for large-scale image recognition. ICLR, 2014.

[31] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmenta-tion. In 3DV 2016, pages 565–571. IEEE, 2016.

[32] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In International Conference on Medical imaging with deep learning, pages 285–296, 2019.

[33] Francesco Caliva, Claudia Iriondo, Alejandro Morales Martinez, Sharmila Majumdar, and Valentina Pedoia. Distance map loss penalty term for semantic segmentation. MIDL, 2019.

[34] Pei Wang and Albert CS Chung. Focal Dice loss and image dilation for brain tumor segmentation. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pages 119–127. Springer, 2018.

[35] Hessam Sokooti, Gorkem Saygili, Ben Glocker, Boudewijn PF Lelieveldt, and Marius Staring. Quantitative error prediction of medical image registration using regression forests. Medical image analysis, 56:110–121, 2019.

[36] Jakub Nalepa, Marek Antoniak, Michal Myller, Pablo Ribalta Lorenzo, and Michal Marcinkiewicz. Towards resource-frugal deep convolutional neural networks for hyperspectral image segmentation. Microprocessors and Microsystems, 73:102994, 2020.

[37] Qing-Song Xu and Yi-Zeng Liang. Monte Carlo cross validation. Chemo-metrics and Intelligent Laboratory Systems, 56(1):1–11, 2001.

[38] Andrew P Bradley. The use of the area under the ROC curve in the eval-uation of machine learning algorithms. Pattern recognition, 30(7):1145–1159, 1997.

· · ·