Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Multiorgan segmentation using distance-aware adversarial networks

Roger Trullo
Caroline Petitjean
Bernard Dubray
Su Ruan

SPIE.

Roger Trullo, Caroline Petitjean, Bernard Dubray, Su Ruan, "Multiorgan segmentation using distance-aware adversarial networks," *J. Med. Imag.* **6**(1), 014001 (2019), doi: 10.1117/1.JMI.6.1.014001.

Journal of Medical Imaging 6(1), 014001 (Jan–Mar 2019)

Multiorgan segmentation using distance-aware adversarial networks

Roger Trullo,[a,*] Caroline Petitjean,[a] Bernard Dubray,[b] and Su Ruan[a]
[a]Normandie University, Institut National des Sciences Appliquées Rouen, LITIS, Rouen, France
[b]Centre Henri Becquerel Normandie Rouen, Rouen, France

Abstract. Segmentation of organs at risk (OAR) in computed tomography (CT) is of vital importance in radiotherapy treatment. This task is time consuming and for some organs, it is very challenging due to low-intensity contrast in CT. We propose a framework to perform the automatic segmentation of multiple OAR: esophagus, heart, trachea, and aorta. Different from previous works using deep learning techniques, we make use of global localization information, based on an original distance map that yields not only the localization of each organ, but also the spatial relationship between them. Instead of segmenting directly the organs, we first generate the localization map by minimizing a reconstruction error within an adversarial framework. This map that includes localization information of all organs is then used to guide the segmentation task in a fully convolutional setting. Experimental results show encouraging performance on CT scans of 60 patients totaling 11,084 slices in comparison with other state-of-the-art methods. [DOI: 10.1117/1.JMI.6.1.014001]

Keywords: segmentation; deep learning; multiorgan; distance map; generative adversarial networks; convolutional neural networks; medical images.

Paper 18142RR received Jun. 30, 2018; accepted for publication Dec. 3, 2018; published online Jan. 10, 2019.

1 Introduction

Computed tomography (CT) is the standard imaging technique used in radiotherapy, a treatment of choice for cancer. The first step for radiotherapy is to identify the target volumes to be treated and the healthy organs to be protected, which are called organs at risk (OAR). In this paper, we focus on the segmentation of thoracic OAR, namely the aorta, heart, trachea, and esophagus, in CT images. As manual segmentation of the OAR is a tedious task and prone to expert variability, automated segmentation would be particularly valuable. These organs have variable shape and size as shown in Fig. 1; in particular, the esophagus is hardly visible. In the following related works section, we will focus on machine learning approaches that, different from purely image-driven methods, make use of external knowledge such as labeled data to guide the segmentation. This has shown to be efficient in noisy or low-contrast environments, as is the case in our application.

1.1 *Related Works in Medical Image Segmentation*

Conventional machine learning algorithms have shown some success in image segmentation. For example in Ref. 1, the authors proposed a patch-based framework with a autocontext model[2] to segment infant brains in MRI images. The association of three-dimensional (3-D) Haar features and random forests allowed obtaining good results. Organ localization is another task, where machine learning algorithms have been applied successfully. The task consists in automatically determining a bounding box of around a particular organ. As an example, in Ref. 3, organ localization in CT scans was performed using regression forests (as opposed to the standard classification random forest). More specifically, a multiclass regression forest was trained to produce six-dimensional probability distributions representing the 3-D bounding boxes of the organs. Moreover, although good performance has been obtained using some of those methods, the results are usually highly dependent on the quality of the registration process and the features representation, and in some cases on both.

Deep learning architectures have outperformed traditional methods and achieved the state of the art in different imaging tasks, such as classification[4,5] and localization.[6] A big advantage of this framework is that the network is able to learn the feature representation automatically from raw images and as such, it avoids the need of engineering complicated features. In classification tasks, for example, the usual framework used is called convolutional neural networks (CNN), which is a succession of several layers conformed by convolutions, a nonlinearity function (like rectified linear units) and pooling operations. The last layers are typically fully connected as in the multilayer perceptron. The output of these networks is a score for each of the classes available in the training set. On the other hand, for semantic segmentation problems, it is required to have one score for each pixel in the image, and although some works have used a classifier for labeling each pixel using a patch as input,[7] it is very expensive computationally due to the large amount of redundant operations (since the patches will be overlapped). Recently in Ref. 8, the authors proposed an architecture called fully convolutional networks (FCN) in which they remove the fully connected layers and change them by convolutional operations, resulting in a network totally composed by convolutions and an architecture that is well suited for semantic segmentation problems as it can produce dense outputs and not only one score per input. FCN have shown great success in semantic segmentation[8] and it improved the state of the art as its first appearance.

*Address all correspondence to Roger Trullo, E-mail: rogertrullo@gmail.com

2329-4302/2019/$25.00

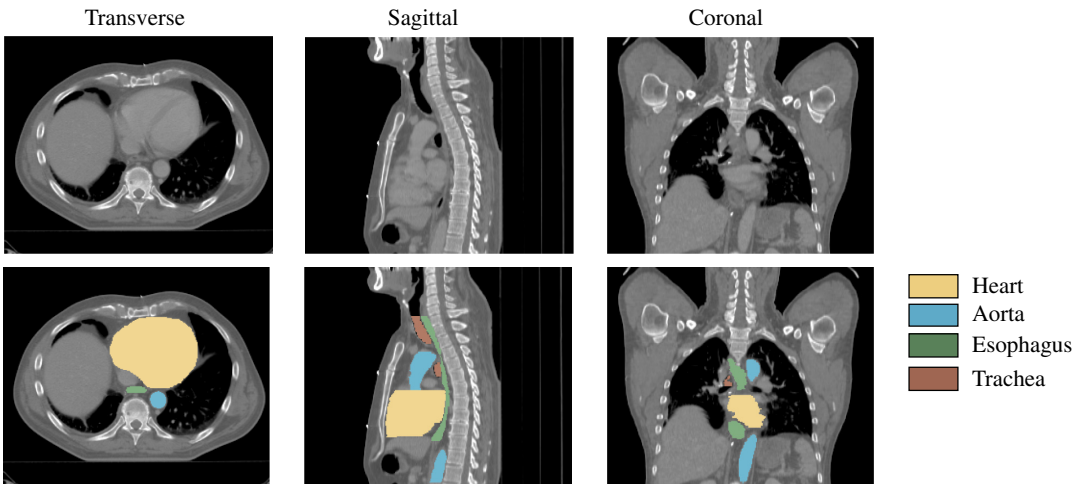

Fig. 1 Typical CT scan with manual segmentation. Some organs like the esophagus are especially challenging to segment.

In the case of medical imaging, several works have adopted deep learning techniques to perform segmentation tasks. Some of them have proposed to use CNNs like in Ref. 9, where a network is trained to take CT patches of a given size as input, and predict whether the central voxel belongs to a lung nodule or not. Similarly, segmentation of OAR has also been treated using CNNs; in Ref. 10, a network is designed to perform the segmentation of several organs in head and neck scans. Still, FCNs remain the most commonly used framework to obtain segmentation maps in medical images. Typically, the network is trained in a patch-to-patch fashion; that is, the output is a segmentation map of the same size of the input. During inference (at testing time), the output patches are stitched together using some fusion method in the overlapping areas where the most common way is majority voting. In the case of MRI images, this framework has been extensively used in brain tissue[11] and brain lesion segmentation.[12] In CT scans, several works have used FCNs to segment organs, such as the liver,[13] esophagus,[14] and even target volumes (tumors).[15] Some specific FCN architectures have shown improved performance on segmentation tasks over others. Notably the U-Net[16] presented an architecture that is able to obtain state-of-the-art performance in the segmentation of neuronal structures in electron microscopic stacks. The architecture is based on the idea of skip connections, first presented in Ref. 8, where the features of shallow layers are combined with the features of deep layers providing a mechanism to overcome the loss of resolution due to the use of pooling operation in deep convolutional networks. As its first appearance, the architecture has been extensively used and has shown top performance in different delineation tasks like pelvic organs in MRI,[17] abdominal organs in CT scans[18] and it can almost be said that it has become the default architecture for segmentation in medical imaging.

1.2 *Our Proposal*

During training, FCNs and variants are typically fed with pairs of images and their associated ground-truth (GT) labels, which in the case of huge datasets, gives cutting edge performance. On the other hand, in medical imaging applications where labeled data can be very expensive to obtain, end-to-end trained, regular architectures may not be enough for the task at hand. To compensate for the low availability of training data, one has to either complexify the learning architecture and/or add external knowledge. In this paper, we present a deep learning-based framework for the automatic segmentation of OARs, based on an FCN network with skip connections, which is enhanced through several procedures. Our approach is motivated by the observation that radiotherapists manually segment the organs based on not only the intensity information, but also the anatomical knowledge they have, e.g., the proximity between organs. The first idea is to learn not only the organ boundaries, but also some localization information: for this, we propose a network that is able to generate the global localization information, represented as a signed distance map. The second idea is to use the generated information as prior spatial context information in an FCN to perform the final segmentation. We show that the performance of the proposed framework is bounded to the quality of the generated localization information for which we show that that an adversarial framework[19] can improve its quality boosting the performance in the final segmentation task.

2 Methods

In the following, we first recall the FCN principle. We explain how to change the loss function when distance maps are considered, and show what questions are raised for multiclass distance maps generation. We then expose the principle of the adversarial networks to produce those maps, and finally, we present our proposed framework.

2.1 *Fully Convolutional Networks for Image Segmentation*

FCNs are a neural network architecture, which is composed totally of convolution operations (unlike CNN that have fully connected layers), which are well suited for dense prediction tasks and are currently the state of the art in many segmentation problems. For medical images, FCN with skip connections is currently the state of the art.[16] In the following, for sake of readability, we will refer to the FCN with skip connections simply as "FCN." Typically they are trained to minimize the cross entropy between the distribution estimated by the network $Q(\hat{y}|x)$ and the GT distribution $P(y|x)$, where y represents the distribution over labels and it is conditioned on the input image x. It can be

shown that minimizing this cross entropy is equivalent to minimize the Kullback–Leibler (KL) divergence between the two distributions. On the other hand, some recent works in medical image segmentation have shown that optimizing a Dice loss function instead of the standard cross entropy loss can lead to a boost in performance[20] in binary segmentation problems as the cross entropy can easily be biased toward the class with greater number of pixels, whereas the Dice loss has an inbuilt normalization. The Dice loss is a soft approximation of the Dice score, which computes the intersection over union ratio of two binary segmentations. The soft approximation is necessary to make the loss function differentiable and instead of using binary values for the output of the network, we use the probability maps:

$$L_{\text{dice}} = \frac{-2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}, \tag{1}$$

where N is the total number of pixels in the image, p_i is the probability of voxel i of being foreground, and g_i is the GT binary value (1 foreground and 0 background) and the minus is used as we use gradient descent for optimization. We can extend the Dice loss to multiclass problems by computing the loss for each organ and adding them up. For an M-class problem, the Dice loss between the probability maps given by the network P and the GT maps G is defined as

$$L_{\text{dice}}(P, G) = -2\sum_{j=1}^{M} \frac{\sum_i^N p_{i,j} g_{i,j}}{\sum_i^N p_{i,j}^2 + \sum_i^N g_{i,j}^2}, \tag{2}$$

where N is the total number of pixels in the image, $p_{i,j}$ is the probability of voxel i of being of class j, $g_{i,j}$ is a binary number equal to 1 if voxel i is of class j, and otherwise is equal to 0. In other words, $g_{i,j}$ is a one hot representation of the GT. Note that the summation is done across organs without taking into account explicitly the background.

It is worth to mention that Eq. (2) could also lead to class imbalance issues. One solution is to add weight parameters that would penalize more errors in smaller organs as shown in Eq. (3), where w_j is the weight for class j. However, in our experiment we did not find these issues and decided to use Eq. (2) instead because it would decrease the number of hyper parameters of the system:

$$L_{\text{dice}}(P, G) = -2\sum_{j=1}^{M} w_j \frac{\sum_i^N p_{i,j} g_{i,j}}{\sum_i^N p_{i,j}^2 + \sum_i^N g_{i,j}^2}. \tag{3}$$

2.2 *Segmentation as a Regression Task*

Instead of representing the label information as one hot vectors, other representations may be more valuable, such as signed distance maps. This representation adds more specific information, such as if the pixel is in the border, which can be particularly important for segmenting foreground adjacent objects. In this case, the GT distance map is computed by assigning to every position the minimum distance (Euclidean in our case) between the position itself and all contour points of the object to segment. By convention, the distance inside an object is chosen to be negative, while outside is positive. In this case, the problem of segmenting objects is treated as a regression task. In particular, a system can be trained to produce distance maps by optimizing its parameters to minimize the difference between its output and the GT distance map. The loss function L_{dmap} in this case can be defined as

$$L_{\text{dmap}}(X, y) = \|y - G(X)\|_2^2, \tag{4}$$

where y is the ground-truth distance map, generated from the binary mask B with $y = \text{distmap}(B)$, and $G(X)$ is the generated distance map from the source image X by a generator network G. After training, one can generate a label map by thresholding the generated distance map at a small value, in theory 0. This approach has proved to work well for the segmentation of small foreground objects that can be very close to each other, as in the extraction of buildings from satellite imagery.[21] Although this approach has been proposed for binary segmentation tasks, it is not clear how to extend the representation to a multiclass setting. Our proposal to handle multiclass distance maps is presented in the following sections.

2.3 *Multiclass Distance Map as Context Information*

In this case, one has M different objects to be segmented in the image. For each object, the GT binary mask B_i is available. One possibility to handle multiclass distance maps would be to generate one map for each class independently and add the losses by extending Eq. (4) to

$$L_{\text{dmap}}(X, Y) = \sum_{i=1}^{M} \|y_i - G(X)_i\|_2^2, \tag{5}$$

where y_i is the ground-truth distance maps, generated from the binary masks B_i, and $G(X)_i$ is the M generated maps from the source image X. However, this model might not be rich enough to take into account context information as the maps would be computed independently, potentially producing uncorrelated outputs.

The spatial information and the relative position of organs are of vital importance for the segmentation, even for human experts performing the task manually. Thus, our proposal is to learn a global prior spatial information that roughly tells the network the relative position of organs. We define this prior global information as a distance map for all the organs; that is, we define a global organ as the union of all the organs and compute a distance map to it:

$$B_{\text{global}} = \bigcup_{i=1}^{M} B_i. \tag{6}$$

Then, the loss of Eq. (4) can be rewritten as

$$L_{\text{dmap}}(X, y) = \|y_{\text{global}} - G(X)\|_2^2, \tag{7}$$

where y_{global} is the distance map obtained from the binary mask B_{global}. In Fig. 2, we show the distance map on the left obtained from its corresponding GT label map on the right. Before explaining our architecture proposal for multiclass segmentation with a global distance map, let us recall the state of the art for image generation tasks.

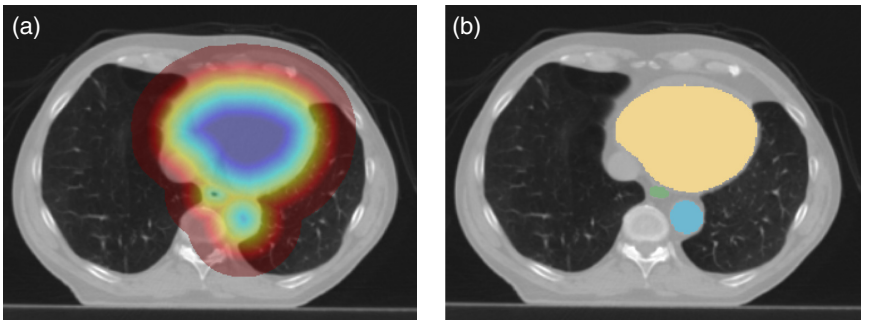

Fig. 2 (a) distance map overlapped on the CT scan and (b) GT labels showing heart in beige, esophagus in green and aorta in blue.

2.3.1 *Generative adversarial networks*

Generative adversarial networks (GANs)[19] are currently the state of the art in imaging generation tasks.[22,23] They are typically used in an unsupervised setting for generative models, where the objective is to produce new images from a distribution of data. It has been used in super-resolution tasks,[23] image completion,[24] and has obtained great attention from the machine learning community. This framework consists of two competing networks such as a discriminator D and a generator G. D is typically a CNN trained to distinguish between real and fake (generated) images, while G is an FCN that is trained to produce realistic images that will try to confuse D. More concretely, the discriminator network D is a CNN that estimates the probability of the input image being drawn from the distribution of real images. That is, D can classify an input image as real or generated. The networks are trained in an alternating fashion, and the idea is that G will generate images that are indistinguishable from those sampled from the real distribution even by a complex discriminator D. In the classical GAN, G takes as input a random vector (typically a sample from a normal distribution) to generate an image. However, a conditional GAN framework[25] can take as input other kind of information to generate images conditioned on the input X. Formally, the loss function for D can be defined as

$$L_D(X,Y) = L_{\mathrm{BCE}}[D(Y),1] + L_{\mathrm{BCE}}\{D[G(X)],0\}, \tag{8}$$

where X is the input image to condition on, Y is the corresponding target image, $G(X)$ is the generated images conditioned on X, and $D(\cdot)$ computes the probability of its input to be real. L_{BCE} represents the binary cross entropy and it is defined as

$$L_{\mathrm{BCE}}(\hat{P},P) = -\sum_i P_i \log(\hat{P}_i) + (1-P_i)\log(1-\hat{P}_i), \tag{9}$$

where P represents the label of the input data and takes its values in $\{0,1\}$ (i.e., 0 for the generated image and 1 for the real target data), and $\hat{P}$ is the predicted probability in $[0,1]$ that the discriminator assigns to the input of being drawn from the distribution of target images.

On the other hand, the loss term used to train G is defined in a way that tries to generate images that are able to "fool" D:

$$L_{\mathrm{ADV}}(X) = L_{\mathrm{BCE}}\{D[G(X)],1\}. \tag{10}$$

Note that GAN has been proposed in a patch-based setting called PatchGAN,[26] with the advantage of having a discriminator that is fully convolutional, making it possible to work with images of arbitrary size while requiring less parameters. In this case, instead of assigning a single value for the whole image, the discriminator assigns one value for every patch of a particular size of the image. To do so, it suffices to use a FCN instead of a CNN, and the loss function [Eq. (8)] is computed for every position in the output and then averaged.

2.3.2 *Generating distance maps with generative adversarial networks*

One can then think of a supervised framework, where the input is X a gray-level CT image and the corresponding GT Y is a distance map. In that way, G would be trained to generate distance maps $G(X)$ conditioned on the input, which is similar to Y enough to confuse the discriminator. The classical reconstruction loss [L2-loss, for example, in Eq. (4)] could be added to the total loss, trying improve the performance in a completely supervised fashion:

$$L_G(X,Y) = L_{\mathrm{dmap}}[G(X),Y] + \lambda L_{\mathrm{ADV}}(X), \tag{11}$$

where λ is a weighting parameter that defines the trade-off between the reconstruction loss and the adversarial loss.

2.3.3 *Image segmentation using distance-aware adversarial networks*

To segment the organs with the help of the distance map, we train a classification FCN using [Eq. (2)], which takes as input not only the input CT image, but also the generated global distance map. The idea is that the network will learn to use the global prior information along with the local CT intensities to improve the performance on the segmentation task.

In total, we propose a two-stage framework as shown in Fig. 3: we first generate the prior global distance map, using an adversarial network, be it GAN or PatchGAN. Then the global distance map is used in a second FCN-based network, to generate the segmentation maps. The framework includes two generators such as the distance map generator and the segmentation generator, which are both based on the same architecture: an FCN with skip connections that are currently the state of the art in segmentation tasks.[16,27,28] The only difference between the distance map generator (Dist. Map Generator) and the segmentation generator network is the number of feature maps in the last layer: there is one feature map corresponding to the global distance

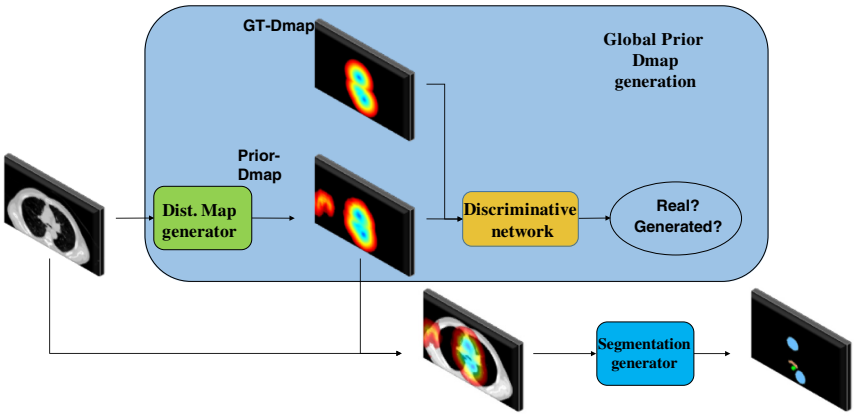

Fig. 3 Proposed framework for multiclass segmentation using distance maps and generative adversarial networks. The details of the generator and discriminator networks are shown in Figs. 4 and 5.

map for the distance map network, and five probability maps corresponding to the four organs and the background, for the segmentation network. The details of the networks G (generative) are shown in Fig. 4, where the numbers indicate the number of feature maps at each layer.

The discriminative network D used in the adversarial part of the distance map generator is shown in Fig. 5. D is a simple FCN as used in the PatchGAN framework.[26] The generator networks use 3×3 kernel with unit stride, whereas the discriminator uses 4×4 kernels with stride of 2; in this way, the resolution of the feature maps is decreased by two at each layer. This in turns augment the receptive field while reducing the size of the feature maps.

Note that the distance map generator is trained with GT distance maps, one question that arises is whether to train the segmentation generator with the GT distance maps or the generated

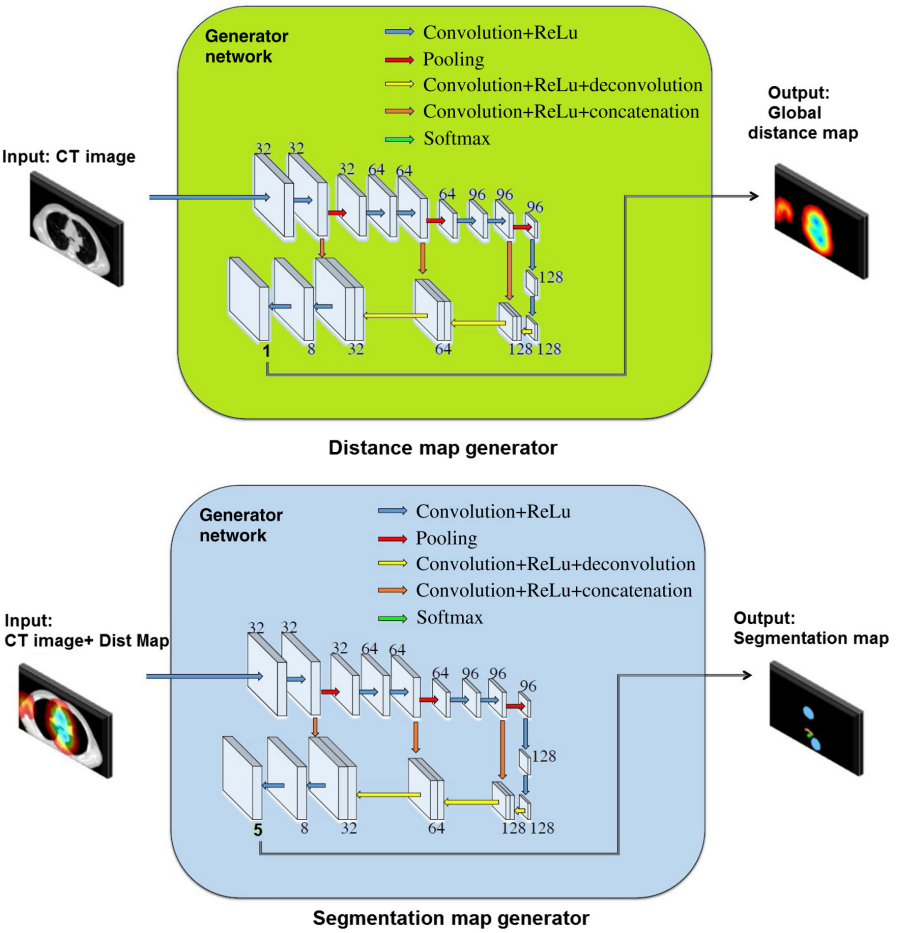

Fig. 4 Details of the generator (G) networks. The difference between the two networks is the number of feature maps in the last layer. The numbers indicate the number of feature maps at each layer.

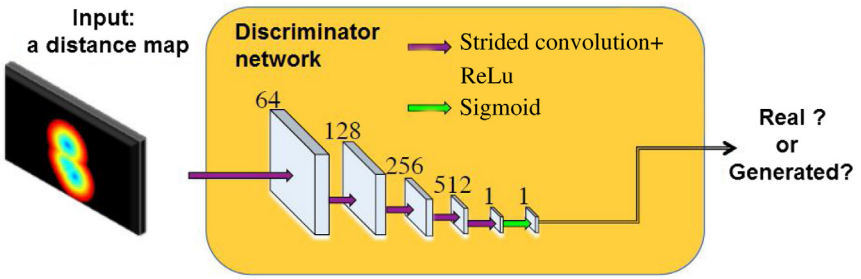

Fig. 5 Details of the discriminator (D) network (PatchGAN setting). The numbers indicate the number of feature maps at each layer.

distance maps of the first network. We will investigate this question in the experimental section.

3 Experiments and Results

3.1 *Dataset and Preprocessing*

CT images (with or without IV contrast) were retrieved from the medical records of 60 patients with nonsmall cell lung cancer referred for curative-intent radiotherapy. On each CT scan, the OARs were delineated by an experienced radiation oncologist. The CT scans have $512 \times 512 \times (150 \text{ to } 284)$ voxels with resolutions in x and y in the range of 0.90 to 1.37 mm and in z in the range of 2 to 3.7 mm; the most frequent resolution is $0.98 \times 0.98 \times 2.5$ mm^3. The intensities in each scan are normalized to have zero mean and unit variance. Dice metric against manually drawn contours is provided in the experiments, using threefold cross validation; that means that the system was trained with 40 patients and evaluated on 20 patients at each fold constructed randomly. Parameter λ in Eq. (11) is set empirically to 0.5.

3.2 *Assessing Adversarial Individual Distance Maps*

We first start by assessing the segmentation based on the generation of individual distance maps of the organs. We compare different ways to generate the distance maps: using a purely generative network (the FCN generator shown in Fig. 4), using a GAN with a classical discriminator, or using a PatchGAN. The generator network is trained using Eq. (11) as loss function. After they are generated, individual distance maps are thresholded, to obtain the organ contours (in practice, the generated distance maps tend to be noisy and we found that typically the areas around the borders of the organs had a small positive distance value instead of zero, thus found a threshold of 1 to work better in practice than the theoretical 0).

Results are shown in Table 1, where dmap + FCN represents the results obtained using only the generator (FCN) without any adversarial term, dmap + GAN refers to the usage of the classical discriminator, and dmap + PatchGAN refers to the use of the PatchGAN discriminator. The results are not particularly good due to independence between organs and the treatment of an intrinsic classification problem (semantic segmentation) as a regression task. This highlights the need for a global distance map, which would gather and consider all organs at once. We found the training of the classical discriminator of the GAN to be rather unstable, whereas the PatchGAN network was more stable, surpassing the regular discriminator results. In the remainder of the experiments, the discriminative network will be the PatchGAN.

Table 1 Segmentation results by learning individual distance maps with an FCN-based generator, a GAN, and a PatchGAN. Distance maps are thresholded to yield segmentations. First row: Dice metric, second row: mean square error between the generated and the GT distance maps. Best results for each organ are in bold.

	dmap + FCN	dmap + GAN	dmap + PatchGAN
Esoph.	0.52 ± 0.13	0.51 ± 0.12	$\mathbf{0.55 \pm 0.13}$
	3.13 ± 0.29	3.28 ± 0.37	$\mathbf{3.06 \pm 0.41}$
Heart	0.62 ± 0.10	0.63 ± 0.09	$\mathbf{0.66 \pm 0.10}$
	37.69 ± 10.47	34.08 ± 11.70	$\mathbf{32.97 \pm 11.27}$
Trach.	0.60 ± 0.11	0.54 ± 0.20	$\mathbf{0.63 \pm 0.10}$
	7.26 ± 0.89	$\mathbf{6.17 \pm 0.81}$	6.93 ± 0.53
Aorta	0.75 ± 0.08	0.70 ± 0.10	$\mathbf{0.76 \pm 0.07}$
	4.38 ± 0.25	$\mathbf{4.02 \pm 0.46}$	4.29 ± 0.49

3.3 *Qualitative Assessment of Global Distance Maps*

In this subsection, we show some qualitative results of the global distance map, which is generated by the first-generator network (Fig. 6). Again, we compare an FCN-based generator without adversarial network, to PatchGAN. The first column shows the distance map obtained by training the FCN generator without any adversarial term, i.e., using only Eq. (4). The second column shows the distance maps obtained using the PatchGAN adversarial framework, i.e., using Eq. (11). Finally, third column shows the distance map obtained from the segmentation GT: we can see that despite some errors, the adversarial framework presents perceptually better results than the FCN. However, both generators will be considered for quantitative assessment of segmentation.

Having these global distance maps, one can now train the segmentation generator using Eq. (2), where the input X is the concatenation of the CT image with distance maps along the channel dimension. We first consider training the

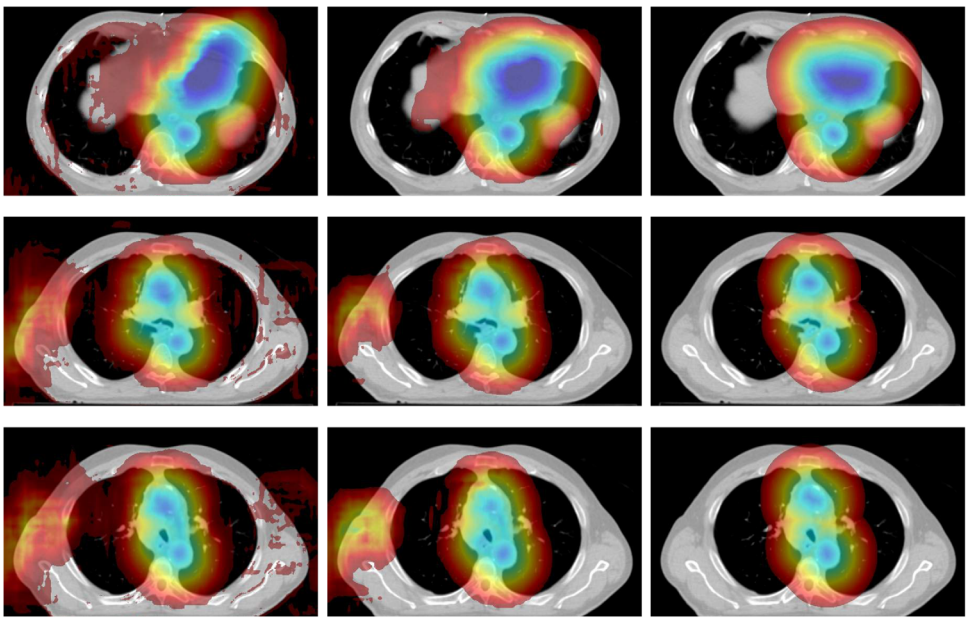

Fig. 6 Distance maps (a) generated by FCN, (b) generated by FCN + PatchGAN, or (c) obtained from segmentation GT.

segmentation network with GT distance maps, and then training the segmentation network with generated distance maps.

3.4 *Training the Segmentation Generator with Ground-Truth Distance Maps*

The segmentation network is trained with GT distance maps. At testing time, we evaluate both the distance maps obtained by the regular FCN (FCN prior), and the ones obtained by the adversarial framework (PatchGAN prior). The results are shown in Table 2, where we also present the performance obtained using the GT distance maps (first column). This case is of course not relevant for real usage as the GT information is unavailable at testing time; it just provides an upper-bound for the system. In accordance with the qualitative results presented just above, the PatchGAN-based prior outperforms the regular, FCN-based distance map prior in all organs except for the heart, for which roughly the accuracy is similar. In the following, we will use, at testing time, the PatchGAN prior.

An issue that arises is that as the segmentation network was trained using GT distance map, at testing time it is not expecting noisy inputs as the ones generated by the distance map generator. Thus, training the segmentation network with the outputs of the first network will not only help the network to learn how to use the global prior information, but also it will make it robust against errors committed in the first stage.

Table 2 Dice segmentation results, using different distance maps during testing of the segmentation generator: GT prior, FCN generated, and PatchGAN generated. The segmentation generator is trained with GT dmaps. Best results for each organ are in bold.

	GT prior	FCN prior	PatchGAN prior
Esoph.	$\mathbf{0.92 \pm 0.04}$	0.48 ± 0.12	0.52 ± 0.13
Heart	$\mathbf{0.98 \pm 0.01}$	0.90 ± 0.06	0.90 ± 0.05
Trach.	$\mathbf{0.95 \pm 0.05}$	0.72 ± 0.14	0.75 ± 0.16
Aorta	$\mathbf{0.95 \pm 0.03}$	0.70 ± 0.16	0.73 ± 0.13

3.5 *Training the Segmentation Generator with Generated Distance Maps*

Now the segmentation network is fed at training time with generated distance maps, output from the first network. In addition, we compare with segmentation results obtained by training the network with distance maps generated by a nonadversarial network and also a regular FCN segmentation network, not making use of prior information (columns "regular FCN" and "FCN + GAN"), its combination with an adversarial framework (FCN + GAN) as have been used in other works.[29,30] In this case, the generator is trained using the same loss function shown in Eq. (11), but exchanging L_{dmap} by L_{dice} is defined in Eq. (2). It is important to note that the label information in this case has to be in a one-hot encoding to be able to compute the different terms in the loss functions. Results are shown in Table 3 and some qualitative results are shown in Fig. 7. When considering the segmentation networks trained without distance maps (two first columns of Table 3), we note that the FCN + GAN approach boosts the performance of the basic FCN; however, we found the training of the FCN + GAN to be rather unstable. Regarding the proposed approach, the results show that (i) the global distance map prior boosts the performance in all the organs except in the trachea (slightly lower); we believe that this lack of improvement is due to its dark intensity that can be confused by background information, specially close to the body; (ii) we can now conclude that the segmentation network should be trained with distance maps generated from the first network, and not GT; (iii) the prior information generated by the adversarial framework (PatchGAN) is better not only perceptually, but also more effective in guiding the segmentation

Table 3 Comparison of Dice segmentation results using segmentation networks trained without distance maps ("No dmaps" column), and with GAN-based generated distance maps. The columns indicate the distance maps used at testing time. Best results for each organ are in bold.

	No dmaps		Segmentation network trained with:		
	Regular FCN	FCN + GAN	GT dmaps	FCN dmaps	PatchGAN dmaps
Esoph.	0.59 ± 0.13	0.66 ± 0.10	0.52 ± 0.13	0.65 ± 0.14	$\mathbf{0.71 \pm 0.15}$
Heart	0.83 ± 0.13	0.85 ± 0.04	0.90 ± 0.05	0.89 ± 0.05	$\mathbf{0.90 \pm 0.06}$
Trach.	$\mathbf{0.84 \pm 0.09}$	0.80 ± 0.09	0.75 ± 0.16	0.80 ± 0.16	0.83 ± 0.10
Aorta	0.82 ± 0.08	0.81 ± 0.12	0.73 ± 0.13	0.82 ± 0.14	$\mathbf{0.85 \pm 0.10}$

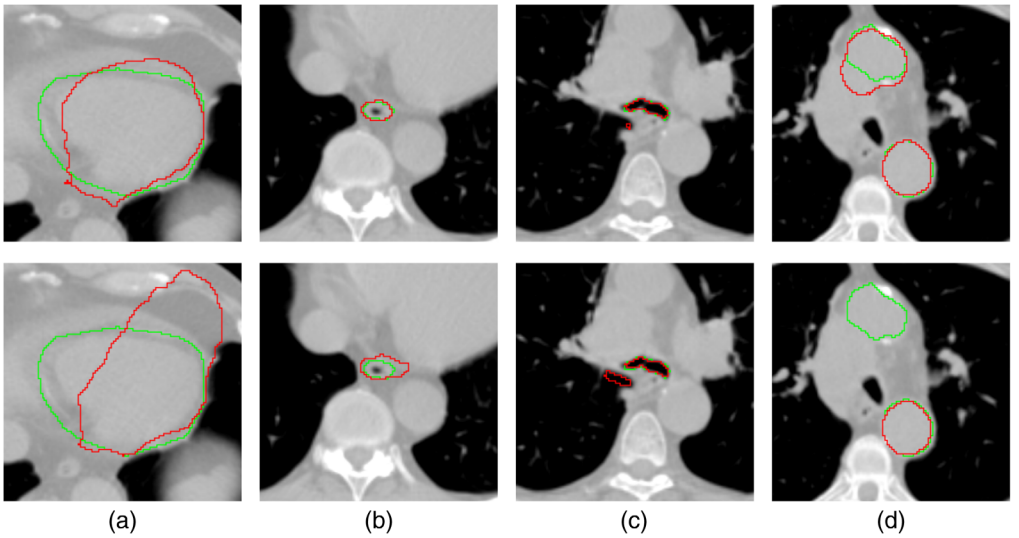

Fig. 7 Segmentation results in red versus GT in green. First row shows the results obtained by the GAN-samples strategy, while second row shows results obtained using the FCN-samples. (a) Esophagus, (b) heart, (c) trachea, and (d) aorta.

network than using GT distance maps, at no additional labeling cost. If we perform a paired two tailed T-test between the Dice results of the proposed framework (PatchGAN dmaps) and the basic FCN (regular FCN), we find the following p-values: 0.01 for the esophagus, 0.03 for the heart, 0.74 for the trachea, and 0.30 for the aorta. By comparing the p-value with the usual threshold of 0.05, the tests suggest that the improvement of the proposed architecture over the basic FCN on the segmentation of the esophagus and the heart is statistically significant. When comparing PatchGAN dmaps and FCN dmaps, all p-values are superior to 0.05; we obtain 0.19 for the esophagus, 0.57 for the heart, 0.48 for the trachea, and 0.44 for the aorta. In this case, the null hypothesis cannot be rejected. Note that the sample size is rather low and a dataset with a larger sample size should be used to state any significant difference.

3.6 *Experiments on a Public Dataset*

With the objective of evaluating the generalization capabilities of our framework, we performed experiments on a public dataset

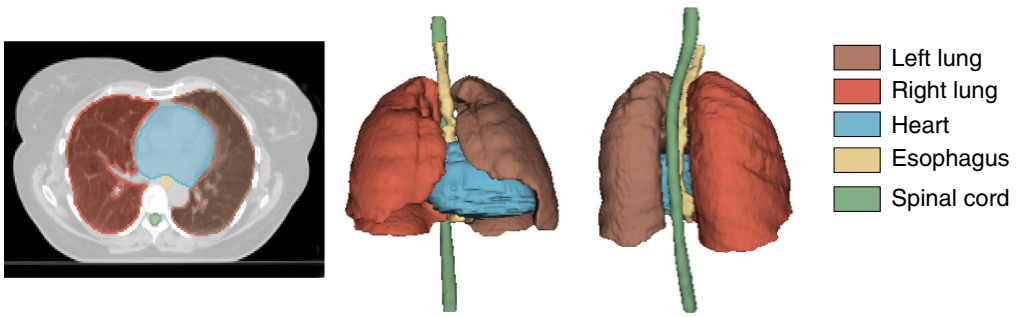

Fig. 8 Data from the AAPM challenge. First column shows a slice of the CT scan with the manual segmentation of the OARs overlapped. The second and third columns show a 3-D rendering of the organs.

that is part of an online challenge. This competition was organized by the American Association of Physicists in Medicine (AAPM) in a live challenge that took place at the AAPM Annual Meeting, in Denver, Colorado, in August 2017.[31,32] The dataset can be found in the website.[31] After the live challenge, submissions are still possible, allowing the evaluation of different algorithms. The challenge consists in the segmentation of five OARs in thoracic CT scans: esophagus, heart, spinal cord, left, and right lung. In Fig. 8, we show a slice with the manual segmentation of these OAR and two different views of the associated 3-D rendering. Annotated scans of 30 patients are available as a training set, whereas the test set includes the scans of 12 patients for which the manual annotation is not available; that means that the evaluation is performed on the challenge website.

3.6.1 *Results*

We performed three experiments showing the effects of our methodology. First, we trained the simple FCN (segmentation map generator in Fig. 4) with the only difference being that the number of channels in the last layer is now 6, representing the five OAR and the background. The data used for training are the 30 patients provided by the challenge and the test set are the 12 subjects evaluated on the AAPM system. We then trained our proposed framework (Fig. 3) with distance maps generated by a regular FCN and by an adversarial framework (FCN dmaps and PatchGAN dmaps, respectively).

From the results shown in Table 4, we can see that the use of prior distance maps does not provide a significant improvement in the lungs and the spinal cord, which have clear contrast in the scan and are thus relatively easy to segment. All methods provide roughly the same Dice metric. On the other hand, the performance of the organs that are more difficult to segment has been improved by the use of the prior context information. This is the case for the heart and the esophagus, whose performances both show a significant improvement: in particular, the p-values obtained after performing a two-tailed T-test for the esophagus are 0.38 for the FCN dmaps and 0.05 for the PatchGAN dmaps, when comparing with the regular FCN. For the heart, the p-values are $3.76e^{-5}$ and $4.03e^{-6}$. Finally, regarding the use of regular distance maps or adversarial distance maps, only the esophagus showed a significant improvement; this suggest that the quality of the context information has a bigger impact in more difficult organs.

Table 4 Comparison of Dice segmentation results using segmentation networks trained without distance maps (No dmaps column), and trained with differently generated distance maps ("segmentation network trained with:" column). Best results for each organ are in bold.

	No dmaps	Segmentation network trained with:	
	Regular FCN	FCN dmaps	PatchGAN dmaps
R. Lung	0.96 ± 0.02	0.96 ± 0.02	$\mathbf{0.97 \pm 0.02}$
Esophagus	0.66 ± 0.10	0.70 ± 0.12	$\mathbf{0.75 \pm 0.11}$
L. Lung	0.96 ± 0.01	0.96 ± 0.01	$\mathbf{0.97 \pm 0.02}$
Heart	0.74 ± 0.08	0.88 ± 0.05	$\mathbf{0.89 \pm 0.03}$
Spinal Crd.	0.88 ± 0.03	0.88 ± 0.03	$\mathbf{0.89 \pm 0.03}$

3.6.2 *Specific evaluation criteria of the American Association of Physicists in Medicine challenge and ranking*

In addition to the Dice index, the AAPM challenge offers two additional metrics upon the submission of the results to the challenge website. First, the mean surface distance (MSD) measures the average distance of two contours. It is based on the directed average Hausdorff measure, which is the average distance of a point in a contour X to its closest point in a contour Y, which is defined as

$$\hat{d}_{\text{avg}}(X,Y) = \frac{1}{|X|}\sum_{x\in X}\min_{y\in Y} d(x,y). \tag{12}$$

However, as this metric is not symmetric, the undirected average Hausdorff measure that is the average of the two directed average Hausdorff measures is used:

$$d_{\text{avg}}(X,Y) = \frac{\hat{d}_{\text{avg}}(X,Y) + \hat{d}_{\text{avg}}(Y,X)}{2}. \tag{13}$$

Second, the directed percent Hausdorff (DPH) measure is also used. For a percentile r, it is defined as the r'th percentile distance over all distances from points in X to their closest point in Y:

$$\hat{d}_{H,r}(X,Y) = K_r[\min_{y\in Y} d(x,y)], \quad \forall\, x \in X, \tag{14}$$

where K_r denotes the r'th percentile. Similarly, the average measure is used:

$$d_{H,r}(X,Y) = \frac{\hat{d}_{H,r}(X,Y) + \hat{d}_{H,r}(Y,X)}{2}. \tag{15}$$

The final normalized score accounts for interobserver variability. In particular, three cases were contoured by multiple observers, and the mean scores of these observers are used as a reference score R against which the submitted contours are compared. For any organ, a perfect value (Dice = 1, MSD = DPH = 0) is defined to have a score of 100. A value equivalent to the average interobserver reference R is given a score of 50. Finally, a score of 0 is given to any result that is below the reference by more than the distance between the perfect score and the reference:

$$\text{Score} = \max\left(50 + \frac{T-R}{P-R} * 50, 0\right), \tag{16}$$

where T is the contour measure (Dice, MSD, or DPH), P is the perfect measure, and R is the interobserver reference measure. The normalized scores for all organs, measure, and test cases, are averaged to give a final score.

Using this evaluation metric, the scores obtained for the basic FCN, FCN dmaps, and PatchGAN dmaps are 40.73, 48.08, and 48.15, respectively. These global scores show a similar trend to the one presented in Table 4, where the use of prior distance maps gives an improvement over the basic FCN and a small performance is gained when using the adversarial distance maps.

3.6.3 *Discussion on the American Association of Physicists in Medicine challenge results*

In this section, we compare the performance of the proposed method to the scores obtained by the seven methods that participated to the challenge.[32] These methods reflect the state of the art in medical image segmentation in anatomical imaging, as five of them are based on CNN, and two of them are based on multiatlas registration and label fusion. One can note the absence of shape-based methods. Many of the CNN-based methods are based on the U-Net architecture,[16] similar to the FCN with skip connections, which is used in this paper. These methods also proceed in a multiscale approach, producing an initial coarse segmentation result, refined in a second step. Differences between the methods include the specific architecture design, the training fashion (from scratch versus fine-tuned), and pre- and postprocessing. As underlined in Ref. 32, distinct analysis of the influence of the network architecture or the training is not possible from the results of the challenge, as both parameters vary together for each participating method. The scores computed as in Eq. (16) and reported in Fig. 9 show that the three top performing methods are CNN-based methods. The proposed method, based on PatchGAN and distance map (diamond on the far right), ranks fourth.

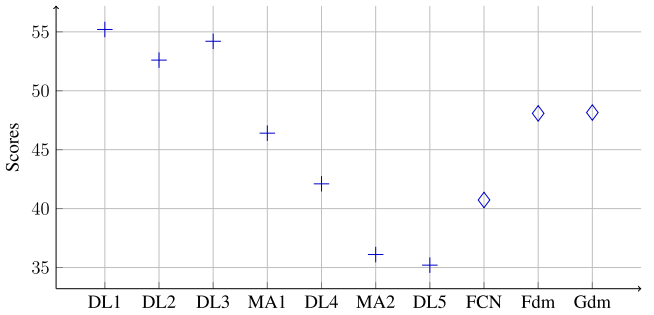

Fig. 9 Scores obtained in the AAPM challenge by the participating methods and by the methods proposed in this paper. Scores from challenge methods (plus symbol) were taken from Ref. 32: DL1-5 refers to deep learning-based methods, MA1-2 refers to multiatlas methods. Scores from the methods presented in this paper (diamond symbol) are FCN, Fdm for FCN-dmaps, and Gdm for PatchGAN-dmaps.

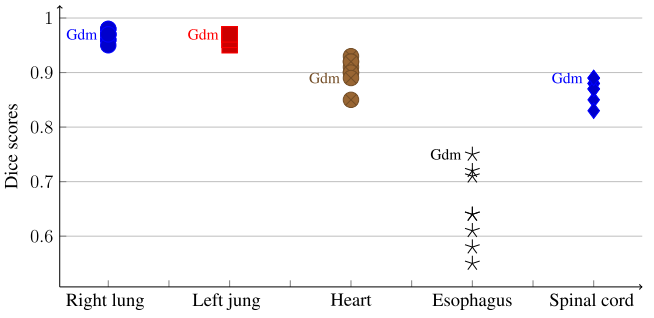

Fig. 10 Dice scores obtained in the AAPM challenge by the participating methods (reported from Ref. 32 and by the PatchGAN-dmaps (Gdm) method proposed in this paper. Each scatter point corresponds to one method, but we omitted the name of the method for the sake of clarity. We only mention where our method stands compared to the others; in particular one can see its Dice scores are better for the esophagus and the spinal cord.

Individual Dice scores for each organ are shown in Fig. 10: one can see that our method obtained the best result for the esophagus, a long and narrow structure difficult to segment, as underlined in the introduction.

4 Conclusions

We presented a framework for the segmentation of four OAR in thoracic CT images: esophagus, heart, trachea, and aorta. It uses state-of-the-art FCN and exploits global localization clues that are used as additional prior information. This additional information is obtained by training a distance map generator network that uses adversarial training to improve the results. We presented several experimental results comparing the distance maps obtained with and without adversarial training, results comparing different variations in our discriminator network; namely regular GAN and PatchGAN, and also showed how this global distance information helps the segmentation network to obtain more accurate results. All the experimental results led to the design choices of our proposed framework, which is able to outperform the FCN in all organs, except the trachea for which similar results are obtained. Finally, although our framework was trained in two steps; one for the localization information and one for the segmentation network, the system is fully differentiable and could in principle be trained in an end-to-end fashion. However, memory requirements become an issue and additional research must be done to reduce the size of the system. Room for improvement is left especially regarding the segmentation of the esophagus. In particular, we used euclidean distance when computing the localization information due to its simplicity; however, this distance has the disadvantage of treating each direction equally without taking into account image information. An option to explore is to find better representation for the localization such as geodesic distances that have been used in different settings in medical image analysis.[33,34]

We have also evaluated our proposed framework on a public dataset, where the objective is to segment other OAR in thoracic CT scans. The results suggest that the usage of prior distance information can be valuable in particular for more challenging organs, such as the heart and esophagus. This is opposed to more simple to segment organs with good contrast like the lungs or the spinal cord. On the other hand, it is important to notice that we used the exact same architecture that we designed for our own dataset with the only difference being the number of classes. One could expect to have an improvement by designing a different architecture.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

This work was cofinanced by the European Union with the European Regional Development Fund (ERDF, HN0002137) and by the Normandie Regional Council via the M2NUM project. The authors' acknowledge the CRIANN (Centre des Ressources Informatiques et Applications Numérique de Normandie, France) for providing computational resources.

References

1. L. Wang et al., "Links: learning-based multi-source integration framework for segmentation of infant brain images," *NeuroImage* **108**, 160–172 (2015).
2. Z. Tu and X. Bai, "Auto-context and its application to high-level vision tasks and 3D brain image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* **32**(10), 1744–1757 (2010).
3. A. Criminisi et al., *Regression Forests for Efficient Anatomy Detection and Localization in CT Studies*, pp. 106–117, Springer, Berlin, Heidelberg (2011).
4. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira et al., Eds., Curran Associates, Inc., pp. 1097–1105 (2012).
5. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," CoRR arXiv:1409.1556 (2014).
6. R. Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, IEEE Computer Society, Washington, DC, pp. 580–587 (2014).
7. D. Ciresan et al., "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in Neural Information Processing Systems 25*, F. Pereira et al., Eds., Curran Associates, Inc., pp. 2843–2851 (2012).
8. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (2015).
9. S. Wang et al., "Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation," *Med. Image Anal.* **40**, 172–183 (2017).
10. I. Bulat and X. Lei, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Med. Phys.* **44**, 547–557 (2016).
11. D. Nie et al., "Fully convolutional networks for multi-modality isointense infant brain image segmentation," in *IEEE 13th Int. Symp. on Biomedical Imaging (ISBI)*, pp. 1342–1345 (2016).
12. K. Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Med. Image Anal.* **36**, 61–78 (2017).
13. Q. Dou et al., "3D deeply supervised network for automatic liver segmentation from CT volumes," *Lect. Notes Comput. Sci.* **9901**, 149–157 (2016).
14. F. Tobias et al., "Esophagus segmentation in CT via 3D fully convolutional neural network and random walk," *Med. Phys.* **44**, 6341–6352 (2017).
15. M. Kuo, D. Jianrong, and L. Yexiong, "Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks," *Med. Phys.* **44**, 6377–6389 (2017).
16. O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," *Lect. Notes Comput. Sci.* **9351**, 234–241 (2015).
17. Q. Zhu et al., "Deeply-supervised CNN for prostate segmentation," in *Int. Joint Conf. on Neural Networks (IJCNN)*, pp. 178–184 (2017).
18. H. R. Roth et al., "Hierarchical 3D fully convolutional networks for multi-organ segmentation," CoRR arXiv:1704.06382 (2017).
19. I. Goodfellow et al., "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680 (2014).
20. F. Milletari, N. Navab, and S. Ahmadi, "V-net: fully convolutional neural networks for volumetric medical image segmentation," in *Fourth Int. Conf. on 3D Vision (3DV)*, pp. 565–571 (2016).
21. J. Yuan, "Learning building extraction in aerial scenes with convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(11), 2793–2798 (2018).
22. A. Radford et al., "Unsupervised representation learning with deep convolutional generative adversarial networks," arXiv:1511.06434 (2015).
23. C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 105–114 (2017).
24. Y. Li et al., "Generative face completion," in *IEEE Conf. on Computer Vision and Pattern Recognition*, (2017).
25. M. Mirza et al., "Conditional generative adversarial nets," CoRR arXiv:1411.1784 (2014).
26. P. Isola et al., "Image-to-image translation with conditional adversarial networks," *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 5967–5976 (2017).
27. P. O. Pinheiro et al., "Learning to refine object segments," in *European Conf. on Computer Vision (ECCV)*, Springer International Publishing, pp. 75–91 (2016).
28. R. Trullo et al., "Joint segmentation of multiple thoracic organs in CT images with two collaborative deep architectures," *Lect. Notes Comput. Sci.* **10553**, 21–29 (2017).
29. P. Luc et al., "Semantic segmentation using adversarial networks," in *NIPS Workshop on Adversarial Training*, Barcelona, Spain (2016).
30. Y. Xue et al., "Segan: adversarial network with multi-scale L_1 loss for medical image segmentation," *Neuroinformatics* **16**, 383–392 (2018).
31. AAPM, "AAPM thoracic auto-segmentation challenge," http://aapmchallenges.cloudapp.net/competitions/3 (13 December 2018).
32. J. Yang et al., "Auto-segmentation for thoracic radiation treatment planning: a grand challenge at AAPM 2017," *Med. Phys.* **45**, 4568–4581 (2018).
33. Z. Wang et al., "Geodesic patch-based segmentation," *Lect. Notes Comput. Sci.* **8673**, 666–673 (2014).
34. G. Wang et al., "DeepIGeoS: a deep interactive geodesic framework for medical image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.* 1–1 (2018).

Roger Trullo received his MS degree from the University of Tours in 2015 and defended his PhD at University of Rouen, France, in 2018. Currently, he is a researcher in Safran, working with machine learning models applied to CT imaging. His research interests include machine learning applied in medical imaging and nondestructive testing.

Caroline Petitjean received her PhD in mathematics and computer science from the University of Paris V, France, in 2003, working on cardiac MR image registration. Since 2005, she has been an associate professor in computer science and signal processing at the University of Rouen. In 2016, she obtained her PhD supervision degree. Her research interests include medical image analysis, segmentation and classification with a focus on statistical shape model and deep learning.

Bernard Dubray is a professor of radiation oncology at the University of Rouen and Centre Henri Becquerel, Rouen, France.

Su Ruan received her MS and PhD degrees in image processing from the University of Rennes, France, in 1989 and 1993, respectively. She is now a full professor at the University of Rouen, France. Her main area of research is image processing, particularly in the fields of image segmentation, pattern recognition, and data fusion. Her developments include advanced machine learning techniques, shape models, graph based image segmentation and data fusion, applicable to medical imaging.