Improving motion-mask segmentation in thoracic CT with multi-planar U-nets

Ludmilla Penarrubia[1,a], Nicolas Pinon[1,b],
Emmanuel Roux[1], Eduardo Enrique Dávila Serrano[1],
Jean-Christophe Richard[1,2], Maciej Orkisz[1], and David Sarrut[1]

[1] *Univ Lyon, Université Claude Bernard Lyon 1, INSA-Lyon, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, LYON, France. and*

[2] *Service de Réanimation Médicale, Hôpital de la Croix Rousse, Hospices Civils de Lyon, France,*

(Dated: November 9, 2021)

[a] Corresponding author: ludmilla.penarrubia@creatis.insa-lyon.fr

[b] Has equally contributed with the first author

Abstract

Purpose. Motion-mask segmentation from thoracic CT images is the process of extracting the region that encompasses lungs and viscera, where large displacements occur during breathing. It has been shown to help image registration between different respiratory phases. This registration step is, for example, useful for radiotherapy planning or calculating local lung ventilation. Knowing the location of motion discontinuity, *i.e.*, sliding motion near the pleura, allows a better control of the registration preventing unrealistic estimates. Nevertheless, existing methods for motion-mask segmentation are not robust enough to be used in clinical routine. This article shows that it is feasible to overcome this lack of robustness by using a lightweight deep-learning approach usable on a standard computer, and this even without data augmentation or advanced model design.

Methods. A convolutional neural-network architecture with three 2D U-nets for the three main orientations (sagittal, coronal, axial) was proposed. Predictions generated by the three U-nets were combined by majority voting to provide a single 3D segmentation of the motion mask. The networks were trained on a database of non-small cell lung cancer 4D CT images of 43 patients. Training and evaluation were done with a K-fold cross-validation strategy. Evaluation was based on a visual grading by two experts according to the appropriateness of the segmented motion mask for the registration task, and on a comparison with motion masks obtained by a baseline method using level sets. A second database (76 CT images of patients with early-stage COVID-19), unseen during training, was used to assess the generalizability of the trained neural network.

Results. The proposed approach outperformed the baseline method in terms of quality and robustness: the success rate increased from 53% to 79% without producing any failure. It also achieved a speed-up factor of 60 with GPU, or 17 with CPU. The memory footprint was low: less than 5 GB GPU RAM for training and less than 1 GB GPU RAM for inference. When evaluated on a dataset with images differing by several characteristics (CT device, pathology, and field of view), the proposed method improved the success rate from 53% to 83%.

Conclusion. With 5-second processing time on a mid-range GPU and success rates around 80%, the proposed approach seems fast and robust enough to be routinely used in clinical practice. The success rate can be further improved by incorporating more diversity in training data *via* data augmentation and additional annotated images from different scanners and diseases. The code and trained model are publicly available.

Keywords: deep learning, segmentation, thoracic CT

I. INTRODUCTION

A. Deformable image registration and motion mask

Various medical applications, such as radiotherapy treatment planning for patients with lung cancer, ventilation assessment in chronic obstructive pulmonary disease (COPD), or recruitment quantification in acute respiratory distress syndrome (ARDS), require aligning lungs and other thoracic structures in 3D CT images, by means of a deformable registration method[1–4]. As deformable image registration is an ill-posed problem, it needs to be regularized and, usually, algorithms are built on an assumption of motion continuity and smoothness, *i.e.*, neighboring points are assumed to have similar displacements[5,6].

This assumption does not hold in the case of thoracic scans representing different phases of the breathing cycle, because lungs and viscera slide along the pleura, so that their displacements are different (generally much larger) than those of the neighboring points of the rib cage. Regularization penalties and smoothness constraints hence lead to erroneous displacement-field estimation around regions of discontinuous motion.

To address this problem, several approaches have been proposed, which consider an initial segmentation of moving and less-moving regions in the image in order to restrict the regularization at the boundaries of the sliding areas[7–11]. As an example, Delmon *et al.*[9] have proposed a B-Spline-based registration method such that regularization is enforced along the direction tangential to the segmented boundary and it is relaxed in the normal direction. Hua *et al.*[11] extended the B-spline approach by incorporating an additional term that acts on a subset of control points, for which the boundary intersects the support of their corresponding basis function. Also a finite-element-based approach has been proposed to perform lung registration preserving motion discontinuity[10]. While such methods often use pre-segmentation of the lungs alone, Vandemeulebroucke *et al.*[12] have developed a so-called *motion mask* that encompasses lungs and viscera, and thus delineates a region, in which the motion field is supposed to be smooth (Figure 1). Its boundary defines where sliding – and thus discontinuity in the vector field – occur between the rib-cage and the internal structures, when breathing. The seminal motion-mask segmentation[12] is based on a level-set framework and involves preliminary automated delineation of several anatomical elements (patient's body contour, lung parenchyma, bony structures) followed by growing an ellipsoid

within the abdomen to fill the thoracic cavity.

This approach, already exploited in previous studies[9,13], has recently been evaluated and clinically used in our institution for radiation therapy of locally advanced non-small cell lung cancer[14]. More than forty patients with 4D thoracic CT, of 10 breathing phases each, were included. Motion masks were automatically extracted for all images ($n > 400$) with the same set of parameters. Average computation time per motion mask was about 12 minutes (including above-mentioned pre-processing steps and level-set algorithm), and up to 16 GB of RAM was required. Each extracted mask was visually assessed and, if needed, manually corrected before use. Approximately half of the automatically segmented masks required such manual correction, which was often time-consuming. Hence, the motion-mask extraction based on level sets has been useful, but is not robust enough to be routinely used in the clinical context.

In the present study, we investigate a more robust and faster motion-mask segmentation method, while considering the level-set approach as a baseline. As image-segmentation methods have dramatically evolved with the advances of deep-learning (DL) algorithms based on neural networks[15], we have chosen to tackle our problem using the DL approach, which – to the best of our knowledge – has not yet been attempted in motion-mask segmentation. The goal here was to assess whether or not a lightweight DL method, trained on acceptable-quality motion masks extracted by level sets, can perform more rapidly and robustly than the baseline method. Therefore, the next section summarizes useful notions from literature on DL-based medical-image segmentation, which guided our choices.

B. Deep-learning segmentation

The general approach to image segmentation by DL techniques is based on the use of three elements: (1) a model devised to produce a segmentation mask given an input image, (2) a dataset composed of images and their associated reference segmentation masks, and (3) an optimisation strategy designed to train the model. The most common models successfully used for medical image segmentation[15] are deep convolutional neural networks based on U-net architecture[16]. This architecture is made of an encoding part and a decoding part. The encoding part projects the input image onto a smaller latent space through several layers of convolution generating so-called feature maps, each followed by a pooling layer to

downscale the resulting feature maps. The decoding part up-scales the projection from latent space to image space, thus predicting a segmentation mask at the same scale as the input image. Skip-connections between the encoding and decoding pathways ensure that details that might be lost during the encoding step can be recovered during the decoding step. Thus, the more abstract concepts encoded in the latent space (multi-scale feature maps) still have access to the finer details of the input image when reconstructing a segmentation mask in the decoding part of the model. The aim of the training procedure is to adjust the parameters of the model so that it can predict which voxels belong to the segmented class, here the area included in the motion mask.

Models based on U-net architecture, which uses 2D convolutions, usually perform very well, but one of their current challenges is to scale them in 3D, *i.e.*, using 3D convolutions with volumetric images as input. Although several teams have proposed successful 3D scaling of the U-net[17–21], each specific application seeks a trade-off between spatial resolution and computational power, since the number of model parameters, as well as the intermediate feature maps have a large impact on memory usage, requiring expensive infrastructures for both training and inference.

In order to tackle the problem with limited computational resources without losing complementary information brought by the 3D context, several approaches have considered extending 2D deep learning models to multi-planar methods (also referred to as 2.5D methods), which take as input several planes extracted from a 3D volume. These approaches can be considered as a sub-category of the multi-stream methods, as described by Litjens *et al.*[15]. In the sequel we focus on the multi-planar methods, as they help reduce the computational-power requirements and improve the applicability of the developed DL methods in clinical routines. Depending on the respective orientation and/or position of the considered planes, three main strategies can be observed: (1) using the classic three orthogonal planes (axial, sagittal, coronal)[22–24], (2) using multiple parallel planes that can form a slab if they are adjacent[25,26] and (3) using multiple planes (usually more than 3) of random orientation[27,28].

Regardless of the orientations/positions, the scope of the extracted input planes has varied across the publications, considering either the whole slice extent[25–27] or a patch (*e.g.*, a sub-part of size 32×32 cropped from a full slice of size 256×256)[22–24,28]. While full slices take into account a larger context, the patch-based approach lightens the computational load, but requires multiple data-augmentation operations[27,28] to compensate for the loss of

context information and improve robustness.

Information from different planes can be merged at several stages of the segmentation pipeline : (1) when the multiple planes are set as multiple channels input to the network[25–28], (2) when passing independently each plane through a network before a fusion layer[22,24] or (3) at the very end of the pipeline with a deterministic ensemble strategy[23]. In (2) and (3), each stream of information passes through a specific branch, and each branch can be a whole network (potentially pre-trained), whereas in (1) there are no branches (only one network).

In the sequel, a multi-planar DL-based strategy for motion-mask segmentation is proposed, as a trade-off between accuracy, computational resources, and speed. Three slightly modified 2D U-nets, each using conventional orthogonal planes (axial, coronal, sagittal) are separately trained, and then merged by majority vote to provide a volumetric binary mask. The results demonstrate the feasibility of this solution, which reasonably compares with a 3D U-net, but other DL architectures may be explored for further improvement.

II. MATERIALS AND METHODS

A. Datasets and expert annotations

Before specifying the method proposed to automatically compute a motion mask from thoracic 3D CT scans, we first describe the data available for training and evaluation.

We used datasets from two clinical trials: ClinicalTrials NCT01635270 and NCT04377685. The first one[14] included 43 patients with locally advanced non-small cell lung cancer treated with radiation therapy. For each patient, a 4D thoracic CT scan composed of ten phases was acquired during free breathing with a Brillance Big Bore (Philips Medical System, Cleveland, OH) and the Pneumo Chest bellows belt used for breathing synchronisation. Images were reconstructed with voxel size ranging from $0.92\times0.92\times2$ mm to $1.37\times1.37\times3$ mm. For three patients, an additional 4D CT acquisition was available. For the current study, we selected from each acquisition 3D CT scans corresponding to the end-exhale and end-inhale phases, *i.e.*, a total of 92 volumes, size $512\times512\times[88-218]$. This dataset will be referred to as S_1. The second dataset, S_2, included 38 early-stage COVID-19 patients, with two breath-hold 3D CT scans acquired at end-exhale and end-inhale on a Siemens Somatom. These scans were of size $512\times512\times[300-400]$ voxels, with voxel sizes ranging from $0.7\times0.7\times1$ mm

to $0.9 \times 0.9 \times 1\,\text{mm}$. For each volume, in both datasets, a motion mask was computed using the baseline level-set method with fixed parameter settings recommended in the seminal publication[12].

A subset of masks from S_1 was used to train the proposed network, while all the data from S_2 were used to evaluate the network's generalizability. To train the network only on correct masks (defined hereafter), two experts in deformable image registration using motion masks (M.O. and D.S.) independently labelled the available S_1 masks segmented by level sets. The correct masks (usable for training) were labelled as either *Full Success* (FS) or *minor Error* (mE), while incorrect masks (not usable for training) were labelled as *Major Error* (ME) or *Full Failure* (FF). This first visual assessment session will be referred to as B1, as each expert was blinded to the other expert's labels.

The experts assigned the FS label to masks perfectly fitting the expected motion discontinuities (*e.g.*, Figure 1), whereas the mE label was assigned when small under- or over-segmentation (*e.g.*, Figure 2a) occurred in non-critical areas, as segmentation errors in regions with small magnitude of lung motion – *e.g.*, near the apex – have less impact on registration. Conversely, ME label was assigned to masks that would require manual editing before use in clinical context, due to their more critical location, large extent (*e.g.*, Figure 2b), inclusion of bones, or exclusion of tumors – or other consolidations – located within the lungs. Eventually, FF label designated completely unusable masks confined in a small portion of the lung (*e.g.*, Figure 2c) or leaking throughout all the volume, generally due to a failure in the initial anatomical segmentation used by the level-sets algorithm. The same criteria were subsequently used to label baseline masks from S_2, as well as the masks segmented by the DL-based methods, as described in Section II C.

After an independent blinded assessment session B1, the two experts jointly adjusted the labels by consensus, session C1, for the cases where initial disagreement had occurred. Thus obtained four-grade labels were subsequently used to split the 43 patients from S_1 into subsets S_1^A (27 patients with at least one correct mask, namely, 14 FS and 35 mE) and S_1^B (16 remaining patients with no correct mask). The former was used for training and validation purposes, as specified in Section II B 3, while the latter made part of the testing subset, as specified in the section II C 1. Data splitting was done on patient basis, as different images of the same patient may not be simultaneously used for training and testing.

B. Segmentation method

1. *Multi-planar U-net framework (majority voting U-nets)*

Our motivation was to associate the high speed, low memory load, and limited parameter number of conventional 2D U-nets with 3D consistency. The proposed approach achieves this goal by merging information from three orthogonal planes, namely, the three (entire) slices from the classic orthogonal directions. The rationale of this approach is that, thanks to the 3D context brought by the orthogonal views, each U-net predicting segmentation in a given slice orientation, *e.g.*, axial, can not only learn a certain regularity of 2D shapes within the slices, but also a regularity along the direction orthogonal to the slices. Predictions performed independently for each slice would not enforce the regularity along the direction orthogonal to the slices.

Hence, the proposed segmentation method is composed of three identical models based on the 2D U-net[16] architecture. The differences with respect to the original U-net are: the number of filters per layer is decreased by a factor of four (to reduce the number of parameters) and batch normalization is performed after each convolution (for training stability). The three U-nets, θ_a, θ_s, and θ_c, are individually trained on 2D slices corresponding to one of the orthogonal planes $v \in \{a, s, c\}$, where a, s, and c respectively stand for *axial, sagittal*, and *coronal*. The set of three trained 2D U-nets will be referred to as $\Theta \equiv \{\theta_a, \theta_c, \theta_s\}$. In the sequel, lower-case letters represent 2D slices and upper-case letters represent 3D volumes. Using a N^3-sized 3D CT scan $X \in \mathbb{R}^{N^3}$ as input, each U-net θ_v is sequentially fed with batches of N^2-pixel slices $x_v^n \in \mathbb{R}^{N^2}, n = 1 \dots N$, along the associated direction v. For each 2D slice, the U-net model θ_v predicts a N^2-sized 2D segmentation mask $\tilde{y}_v^n$. The training process adjusts the weights of θ_v so as to minimize the dissemblance between $\tilde{y}_v^n$ and the corresponding reference mask $\overset{*}{y}{}_v^n$, measured by a loss function $L(\tilde{y}_v, \overset{*}{y}{}_v)$ based on the Dice similarity coefficient (see section II B 3). In the experiments, $N = 256$ was used in agreement with data resampling strategy detailed in the section II B 2.

In the inference phase, three 3D segmentation masks $\tilde{Y}_a$, $\tilde{Y}_s$, and $\tilde{Y}_c$ are built from the respective θ_v U-net outputs by concatenating the obtained 2D masks $\tilde{y}_v^n, n = 1 \dots N$. These three masks are eventually combined to compute the final prediction $\tilde{Y}_f$ with 3D consistency. They can be merged in different ways: union, intersection or majority vote. While union

and intersection could respectively lead to over- or under-segmentation, majority vote allows one mask to fail without downgrading the final results in the areas where the other two are successful. Hence we have chosen the majority vote strategy, *i.e.*, a voxel located at (i, j, k) in $\tilde{Y}_f$ is set to one if at least two out of the three U-nets have predicted one for this location:

$$\tilde{Y}_f(i,j,k) = \begin{cases} 1 \text{ if } \sum_v \tilde{Y}_v(i,j,k) \geq 2 \\ 0 \text{ otherwise.} \end{cases} \quad (1)$$

2. Pre-processing

All images were resampled and resized to obtain isotropic volumes of $256 \times 256 \times 256$ voxels with 2 mm resolution consistent with the average slice spacing in the dataset S_1. Missing axial slices, if any, were padded with -1000 HU value corresponding to the air.

3. Training

Considering the number of annotated data usable for training (27 patients in S_1^A), we applied K-fold cross-validation scheme ($K = 9$) to train the model with as many images as possible, while leaving out a subset of annotated data for final evaluation. Indeed, using $K = 9$ allowed us to equally split S_1^A into groups of $27/9 = 3$ patients. Hence, in each of the 9 folds, data from $7 \times 3 = 21$ patients, were used for actual training, data from 3 distinct patients were used for validation, and data from 3 other patients were left out for testing.

The models Θ were trained during 20 epochs with a fixed batch size of 32 slices. Their parameters (weights) were updated by the Adam optimizer[29] with 1×10^{-3} learning rate that was fixed after a grid search with tested values $\{1 \times 10^{-3}, 5 \times 10^{-4}, 1 \times 10^{-4}, 5 \times 10^{-5}, 1 \times 10^{-5}\}$ for the learning rate and $\{32, 64, 128\}$ for the batch size. The loss function was $L = 1 - D$, where D stands for the Dice similarity coefficient frequently used to measure the overlap between segmentation results[30]; its values range from zero (no overlap) and one (perfect overlap).

C. Evaluation method

1. Models and respective test sets

With the 9-fold cross-validation strategy the segmentation predicted by the proposed method was evaluated exactly once for each image from the subset S_1^A containing mostly correct baseline masks. To also evaluate the predictions on images where the baseline segmentation failed, we complemented each testing set by two patients drawn from S_1^B, so that each segmentation predicted for this subset was also evaluated exactly once. Thus, within the data used in a given fold, the training, validation, and testing sets respectively represented around 70%, 10%, and 20%, and each scan from the testing set was segmented by Θ trained on 21 patients.

Subsequently, we trained Θ on all images from the S_1^A subset with baseline masks annotated as correct. Thus trained models, referred to as Θ^{all}, were used in three additional experiments. In the first experiment Θ^{all} were used to segment all images from S_1^B and thus assess to what extent the performance of the proposed framework is affected by increasing the training set (from 21 to 27 patients). The second experiment aimed to evaluate the generalizability of the proposed segmentation framework by applying Θ^{all} to all images from S_2.

The last experiment aimed at comparing the proposed approach with a 3D U-net[31]. The 3D U-net was also trained on all images from the S_1^A subset with baseline masks annotated as correct, and then applied onto all images from S_1^B. For a fair comparison, the same hardware was used for training and inference, and the hyper-parameters were set accordingly: same image size and resolution (256^3 voxels, 2 mm resolution), batch size of 1 (limited by memory constraint) compensated by an increased number of epochs (160) to ensure an equal total number of iterations.

2. Evaluation process

In the absence of ground truth segmentation for all the considered testing data, the evaluation was mainly based on labels assigned by the two experts according to the criteria described in Section II A. Similarly to the visual-assessment sessions B1 (blinded) and C1 (consensus) for baseline motion masks alone, additional sessions were performed for both

baseline and predicted masks, as follows. During a second blinded session (B2) each observer independently labelled the masks segmented by both methods, baseline and proposed, in all images from the set S_1. Labels assigned during B2 to the baseline masks were compared to those assigned during B1, so as to assess each expert's intra-observer variability by means of the Cohen's kappa coefficient[32], while labels independently assigned by different observers within the same session were used to assess the inter-observer variability. The session B2 was followed by a consensus session (C2), during which the observers agreed on initially discordant labels for the masks predicted by the proposed method. Thus obtained consensus labels were unique for all images from the set S_1 and were compared to the consensus labels assigned to the baseline masks during C1. In each session, the masks were presented in a random order and the observer was not informed by which method the mask was segmented. The observer freely scrolled the axial, sagittal, and coronal slices of the original CT scan with an adjustable-density translucent mask superimposed onto the image gray levels.

The same procedure was followed during two additional sessions, blinded (B3) and consensual (C3), to label masks segmented by both methods in the images from the dataset S_2. Eventually, the masks segmented in the images from the S_1^B subset by the proposed method using Θ^{all} and by the 3D U-net were labelled according to the same procedure during a blinded session (B4) and a consensus session (C4).

In addition to the above-described semi-quantitative (categorical) comparisons, a quantitative comparison was carried out to assess the benefits of the majority voting. In the absence of ground truth, we used as reference the subset of CT scans from S_1^A, for which the motion mask segmented by the baseline method was labelled as full success (consensus label FS, $n = 14$). We used the Dice score, the average symmetric surface distance (ASSD), and the Hausdorff distance measuring the largest gap between the surfaces. These measures were calculated for the masks $\tilde{Y}_f$ predicted by majority voting and for the masks $\tilde{Y}_a$, $\tilde{Y}_s$, and $\tilde{Y}_c$ separately predicted by each of the 2D U-nets.

III. RESULTS

We first present the results of the proposed method compiled from nine testing subsets of the 9-fold cross-validation, evaluated on S_1. Then we describe the results of the DL-based methods obtained using the models (Θ^{all} and 3D U-net) trained on all patients from S_1^A and

evaluated on S_1^B; we also report the generalizability of Θ^{all} evaluated on S_2. Eventually, experts' evaluation variability, as well as memory and time requirements are reported.

A. Nine-fold cross-validation

Figure 4 displays the evolution of the training loss function (cyan curve) for the θ_s U-net. The validation loss was evaluated at the beginning of the training process and at the end of each epoch (blue dots). The fact that the values of the training and evaluation losses are very close indicates that the proposed model did not lead to overfitting on the training data. A fast decrease during the two first epochs and stabilization around the fifth epoch can be observed. The same behavior was observed for θ_a and θ_c, their corresponding graphs as well as more details can be found in the Supplementary Material (Section Training loss graphs, Figures S.1 and S.2).

The Table I represents the confusion matrix allowing a semi-quantitative (categorical) comparison between the consensus labels (session C1 vs. session C2) assigned to the motion masks segmented from the S_1 dataset by the baseline level-set algorithm (columns) and by the proposed method (rows). The proposed method yielded 79% of correct masks (21% of FS labels and 58% mE labels) against only 53% (15% FS and 38% mE) for the baseline algorithm. Green color highlights 44 improvements (48%), *i.e.*, images for which the predicted motion mask received a better label than the baseline one; 33 out of them had incorrect (unusable for registration, *i.e.*, labels FF or ME) baseline masks, while correct masks were predicted by the proposed method. Conversely, 16 predicted masks received a worse label than the baseline one (17%, red color). Nine of them passed from correct to incorrect category. Labels remained unchanged for 32 images (35%). There was no FF in the predicted masks, while the baseline method failed in 17 CT scans. Among these, the proposed method predicted 13 masks considered as usable for registration (2 FS and 11 mE).

Figure 5 shows side-by-side orthogonal slices from motion masks predicted by each individual U-net θ_v and by majority vote. It can be seen that each individual U-net predicted consistent segmentation in the planes in which it was specialized, but inconsistencies (holes or disconnected extra regions) can be observed in the remaining planes. The multi-planar strategy, merging by majority voting the segmentation produced by three orthogonal U-nets, allowed the correction of these inconsistencies. Table II confirms the improvement resulting

from the majority voting as compared to the predictions made by each 2D U-net separately. It also shows an overall good agreement – in terms of overlap and average surface distance – between the predicted motion masks and those obtained by the baseline method and labelled as full success. A relatively high Hausdorff distance corresponds to localized small outliers. More details can be found in the Supplementary Material (Section Quantitative comparison between majority vote and each 2D U-net, Figures S.5 through S.7).

B. Models trained on all available correct masks

1. *Evaluation on the subset where the baseline method failed*

The training curves of the models Θ^{all} and 3D U-net are provided in the Supplementary Material (Section Training loss graphs, Figures S.3 and S.4, respectively). Let us remind that these models were evaluated on the subset S_1^B (36 CT scans), on which the baseline method failed (no correct label). In comparison with these baseline labels, Θ^{all} performed better on 34 scans (94.4%), of which 33 changed the labels from incorrect to correct, and the 2 remaining labels were unchanged (ME).

Compared with the masks obtained for the same subset (S_1^B) during the testing phase of the 9-fold cross-validation (Section III A), the results obtained with Θ^{all} were improved in 9 scans (25%), unchanged in 26 scans (72.2%), and degraded in 1 scan (2.8%) from FS to mE. This result shows the improvement brought by increasing the training set (from 21 to 27 patients).

Among the labels obtained by the 3D U-net, 31 (86.1%) were improved and 5 (13.9%) remained unchanged (ME) in comparison with the baseline; for 24 scans (66.7%) the label changed from incorrect to correct. Details can be found in the Supplementary Material (Section Proposed method Vs. 3D U-net, Figure S.8).

2. *Generalizability*

Table III summarizes the semi-quantitative results (session C3) evaluating the generalizability of Θ^{all} when applied on S_2 (different disease, different hospital). The only full failure was produced by the baseline method, which was successful in 40 CT scans (53%). The proposed method obtained an improvement for 34 scans (of which 25 passed from incorrect

to correct category) and a deterioration for 4 scans (of which 2 became incorrect), so that 83% of the masks predicted by the proposed method were correct. Masks obtained with both methods received the same labels in 50% of scans (38 masks).

C. Experts' variability, memory and time requirements

To assess the agreement between the experts' annotations, the intra-observer variability was computed between B1 and B2 evaluation sessions on the masks produced by the baseline method ($n = 92$). This resulted in a Cohen's kappa of 0.69 (confidence interval[32] $[0.54, 0.84]$) and 0.66 (CI $[0.50, 0.83]$) for Experts 1 and 2, respectively. As for the inter-observer variability, it was computed with the evaluations by both experts on the union of labels from B1, B2 , B3, and B4 sessions on the masks produced by the baseline and proposed methods ($n = 500$). The result was a Cohen's kappa of 0.54 (CI $[0.45, 0.62]$). The confusion matrices underlying the computation of Cohen's kappa are provided in the Supplementary Material (Tables S.I through S.III).

The training of Θ was performed with a NVIDIA Tesla V100 GPU, using approximately 3.7 GB of graphic memory and 2.6 GB of RAM (vs. 9 GB of graphic memory and 3.5 GB of RAM to train the 3D U-net). One fold training for each model (θ_a, θ_s, and θ_c) took about 15 minutes. The overall training of all three models on the 9 folds (21 patients per fold) took almost 7 hours. The training of Θ^{all} (on 27 patients) required $45 \times 3 = 135$ minutes (2 hours 15 minutes), while the training of the 3D U-net on the same data took 62 hours. For the inference, computing a 3D motion mask with the proposed method required 0.8 GB of graphic memory and was completed within 5 seconds using a Quadro P2000 GPU or 35 seconds using only the CPU (Intel©Core™ i5-8500 CPU @ 3.00 GHz x 6). The storage of each model θ_v weights required 13 MB, *i.e.*, 39 MB for the entire model Θ. With the 3D U-net, the inference was completed within 6 seconds on GPU and 10 seconds on CPU and required 4 GB of graphic memory, while the storage of the model weights needed 25 MB.

IV. DISCUSSION

The goal of the work herein presented was to assess whether or not the motion-mask segmentation can be achieved by a relatively lightweight DL method more rapidly and

robustly than with the baseline level-set method. To this purpose, we placed ourselves in an application-driven perspective: mid-range GPU for training, no manual expert-made segmentation available for training, no data augmentation nor heavy model design, so as to fit the inference-time and memory constraints required by a usage in clinical routines (even without GPU available).

A multi-planar U-net framework was proposed to automatically segment motion masks from thoracic CT scans. It was compared with a standard 3D U-net on a subset of the available data. Overall, both DL-based solutions received similar scores and outperformed the baseline level-set method[12]. Specifically for the proposed multi-planar framework, combining predictions of different U-nets by majority vote was shown to be beneficial compared to the separate U-net predictions (see Figure 5 and Table II). The proposed solution was more robust than the baseline level-set method[12], when applied to unseen data from S_1: correct motion masks represented 79% of the masks predicted *vs.* 53% for the baseline method, and the proposed solution produced no full-failure mask *vs.* 19% for the baseline method. In the experiments conducted on the dataset S_2 (unlike S_1, no subset of S_2 was used for training), the proposed method also outperformed the baseline one, with respectively 83% *vs.* 53% of correct motion masks. Let us emphasize that the two datasets, S_1 and S_2, were different in terms of disease analyzed (lung cancer *vs.* COVID), scanner (Philips *vs.* Siemens), and acquisition protocol, *e.g.*: 4D (S_1) versus dual breath-hold (S_2), larger field-of-view for S_1 compared to S_2.

These comparisons were based on labels assigned by two experts whose average intra-observer agreement assessed by the Cohen's kappa coefficient was in the range 0.60 - 0.79, which means that both experts were moderately consistent[32] in their individual evaluations. Their inter-observer agreement before consensus was relatively weak (0.52), which justified the use of the consensus step.

In terms of memory, both DL-based solutions are sufficiently lightweight to segment a $256 \times 256 \times 256$ voxel image on a standard computer with a mid-range GPU board (5GB of graphic memory) or even without it, but the proposed multi-planar U-net required five times less graphic memory than the 3D U-net. The inference times of both DL-based solutions were very similar to each other and outperformed the baseline method ($\geq$ 120 times and $\geq$ 20 times faster using GPUs and CPUs, respectively).

Also, the baseline method[12] needed three masks as inputs: lung mask, rough bony

anatomy and patient's body outline. Although the two latter are relatively simple to provide, the lung mask can be more difficult to obtain in presence of dense regions within the lungs. The DL-based methods do not need any pre-segmentation as input and are self-contained.

The work herein presented can be considered as a proof of concept. To the best of our knowledge, there is no publication reporting the use of DL models for motion-mask segmentation. Our goal was to make a step forward with respect to the reference method, based on level sets and representing the state-of-the art, rather than to seek the best-performing DL architecture. Interested teams can propose improvements with respect to thus established benchmark (part of the training datasets can be accessed upon request). The improvement potential of the proposed multi-planar method can also be foreseen within the same framework. First, increasing the number and accuracy of annotated data should enhance the robustness of the trained model. In the present work, only 49 correct masks were available in total – and split between training, validation, and testing – and only 14 of these could be considered as actual reference (full success), while the remaining 35 contained minor errors. The model should be re-trained upon availability of more reference masks carefully drawn or corrected by experts. Adding annotated data from S_2 and from other datasets, upon their availability, should reinforce the robustness of the model, as demonstrated by the noticeable improvement in performance when increasing the training set from 21 to 27 patients (Section III B). In the context of enriching the models using newly expert-annotated images, the proposed multi-planar U-nets may be preferred over the 3D U-net, as the former allows retraining the model within a few hours vs. several days with the latter. Second, data-augmentation techniques can also be used: both standard (linear and non-linear image transformations) and specific (*e.g.*, simulating local condensations within the lungs). Also, many small outliers responsible for large Hausdorff distances might be cleaned by simple post-processing techniques such as retaining the largest connected components, hole filling, and smoothing. Finally, there are also avenues that can be explored to improve the framework itself, such as replacing the majority voting by a learned-merging strategy[22,24].

A limitation of our evaluation was the absence of reference motion masks segmented by experts, so that quantitative comparisons (overlap and surface distances) could be carried out only on a subset of data for which the masks segmented by the baseline method were considered fully successful. Thus obtained measures were biased, because it was impossible to obtain them for the cases where the proposed method performed (visually) better than

the baseline one.

Nevertheless, the proposed approach has already shown more robust than the baseline method and outperformed it in terms of computational cost. It performed comparably to a 3D U-net in terms of robustness, while requiring less memory and being much faster to train. The code and weights of the proposed model as well as a practical example for applying motion-mask segmentation are available to the community at: https://github.com/emmanuelrouxfr/deep_learning_motion_mask_segmentation.

ACKNOWLEDGEMENTS

This work was performed within the framework of the LABEX PRIMES (ANR-11-LABX-0063) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR), and the SIRIC LYriCAN Grant INCa-INSERM-DGOS-12563. This work was granted access to the HPC resources of IDRIS under the allocation 2019-101203 made by GENCI (Jean Zay computing center).

CONFLICT OF INTEREST

The authors have no conflicts of interest to disclose.

DATA AVAILABILITY STATEMENT

S_1 dataset that supports the findings of this study is available on request from the corresponding author. These data are not publicly available due to privacy or ethical restrictions. S_2 dataset that supports the findings of this study is available from CHU of Saint-Etienne/CREATIS Laboratory. Restrictions apply to the availability of these data, which were used under license for this study. Data are available at : https://covid.creatis.insa-lyon.fr with the permission of CHU of Saint-Etienne/CREATIS Laboratory.

REFERENCES

[1] J. R. McClelland, M. Modat, S. Arridge, H. Grimes, D. D'Souza, D. Thomas, D. O. Connell, D. A. Low, E. Kaza, D. J. Collins, M. O. Leach, and D. J. Hawkes, "A generalized framework unifying image registration and respiratory motion models and incorporating image reconstruction, for partial image data or full images," Physics in Medicine and Biology **62**, 4273–4292 (June 2017).

[2] C. Guy, E. Weiss, G. Christensen, N. Jan, and G. Hugo, "CALIPER: A deformable image registration algorithm for large geometric changes during radiotherapy for locally advanced non-small cell lung cancer," Medical Physics **45**, 2498–2508 (2018).

[3] K. J. Chae, J. Choi, G. Y. Jin, E. A. Hoffman, A. T. Laroia, M. Park, and C. H. Lee, "Relative regional air volume change maps at the acinar scale reflect variable ventilation in low lung attenuation of COPD patients," Academic Radiology **27**, 1540–1548 (2020).

[4] M. Orkisz, A. Morales Pinzón, J.-C. Richard, C. Guérin, L. Solórzano Vargas, D. Sicaru, C. García-Hernández, M. Gómez-Ballén, B. Neyran, E. Dávila Serrano, and M. Hernández Hoyos, "Voxel-wise assessment of lung aeration changes on CT images using image registration: application to acute respiratory distress syndrome (ARDS)," International Journal of Computer Assisted Radiology and Surgery **14**, 1945–1953 (2019).

[5] A. Sotiras, C. Davatzikos, and N. Paragios, "Deformable Medical Image Registration: A Survey," IEEE Transactions on Medical Imaging **32**, 1153–1190 (July 2013).

[6] D. Sarrut, T. Baudier, M. Ayadi, R. Tanguy, and S. Rit, "Deformable image registration applied to lung SBRT: Usefulness and limitations," Physica Medica **44**, 108–112 (Dec. 2017).

[7] Z. Wu, E. Rietzel, V. Boldea, D. Sarrut, and G. C. Sharp, "Evaluation of deformable registration of patient lung 4DCT with subanatomical region segmentations: Evaluation of deformable registration of 4DCT with segmentations," Medical Physics **35**, 775–781 (Jan. 2008).

[8] A. Schmidt-Richberg, J. Ehrhardt, R. Werner, and H. Handels, "Slipping Objects in Image Registration: Improved Motion Field Estimation with Direction-Dependent Regularization," Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009, **LNCS 5761**, (London, UK), 755–762, Springer, Berlin, Heidelberg (2009).

[9] V. Delmon, S. Rit, R. Pinho, and D. Sarrut, "Registration of sliding objects using direction dependent B-splines decomposition," Physics in Medicine and Biology **58**, 1303–1314 (2013).

[10] A. Derksen, S. Heldmann, T. Polzin, and B. Berkels, "Image registration with sliding motion constraints for 4D CT motion correction," Informatik Aktuell – Bildverarbeitung für die Medizin 2015, (Lübeck, Germany), 335–340, Springer, Vieweg, Berlin, Heidelberg (2015).

[11] R. Hua, J. M. Pozo, Z. A. Taylor, and A. F. Frangi, "Multiresolution eXtended Free-Form Deformations (XFFD) for non-rigid registration with discontinuous transforms," Medical Image Analysis **36**, 113–122 (2017).

[12] J. Vandemeulebroucke, O. Bernard, S. Rit, J. Kybic, P. Clarysse, and D. Sarrut, "Automated segmentation of a motion mask to preserve sliding motion in deformable registration of thoracic CT," Medical Physics **39**, 1006–1015 (2012).

[13] A. Morales Pinzón, M. Orkisz, J.-C. Richard, and M. Hernández Hoyos, "Lung segmentation by cascade registration," IRBM **38**, 266 – 280 (2017).

[14] M. Ayadi, T. Baudier, G. Bouilhol, P. Dupuis, P. Boissard, R. Pinho, A. Krason, S. Rit, L. Claude, and D. Sarrut, "Mid-position treatment strategy for locally advanced lung cancer: A dosimetric study," The British Journal of Radiology **93**, 20190692 (Apr. 2020).

[15] G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. Adiyoso Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," Medical Image Analysis **42**, 60 – 88 (2017).

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, **LNCS 9351**, (Munich, Germany), 234–241, Springer, Cham (2015).

[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-net: Learning dense volumetric segmentation from sparse annotation," Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, **LNCS 9901**, (Athens, Greece), 424–432, Springer, Cham (2016).

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

[19] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, and P.-A. Heng, "H-DenseUNet: Hybrid densely connected unet for liver and liver tumor segmentation from CT volumes," IEEE Transactions

on Medical Imaging **37**, 2663–2674 (2018).

[20] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. J. Wirkert, and K. H. Maier-Hein, "nnU-net: Self-adapting framework for U-net-based medical image segmentation," Informatik aktuell – Bildverarbeitung für die Medizin (Lübeck, Germany), 22, Springer, Vieweg, Wiesbaden (2018).

[21] S. E. Gerard, J. Herrmann, D. W. Kaczka, G. Musch, A. Fernandez-Bustamante, and J. M. Reinhardt, "Multi-resolution convolutional neural networks for fully automated segmentation of acutely injured lungs in multiple species," Medical Image Analysis **60**, 101592 (2020).

[22] P. Moeskops, J. M. Wolterink, B. H. Van der Velden, K. G. Gilhuijs, T. Leiner, M. A. Viergever, and I. Išgum, "Deep learning for multi-task medical image segmentation in multiple modalities," Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, **LNCS 9901**, (Athens, Greece), 478–486, Springer, Cham (2016).

[23] M. Lyksborg, O. Puonti, M. Agn, and R. Larsen, "An ensemble of 2D convolutional neural networks for tumor segmentation," Scandinavian Conference on Image Analysis – SCIA 2015, **LNCS 9127**, (Copenhagen, Denmark), 201–211, Springer, Cham (2015).

[24] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013, **LNCS 8150**, (Nagoya, Japan), 246–253, Springer, Berlin, Heidelberg (2013).

[25] X. Han, "Automatic liver lesion segmentation using A deep convolutional neural network method," CoRR **abs/1704.07239** (2017).

[26] G. Wardhana, H. Naghibi, B. Sirmacek, and M. Abayazid, "Toward reliable automatic liver and tumor segmentation using convolutional neural network based on 2.5d models," International Journal of Computer Assisted Radiology and Surgery **16**, 41–51 (2020).

[27] M. Perslev, E. B. Dam, A. Pai, and C. Igel, "One network to segment them all: A general, lightweight system for accurate 3D medical image segmentation," Medical Image Computing and Computer Assisted Intervention – MICCAI 2019, **LNCS 11765**, (Shenzhen, China), 30–38, Springer, Cham (2019).

[28] H. R. Roth, L. Lu, A. Seff, K. M. Cherry, J. Hoffman, S. Wang, J. Liu, E. Turkbey, and R. M. Summers, "A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations," Medical Image Computing and Computer-Assisted

Intervention – MICCAI 2014, **LNCS 8673**, (Boston, MA, USA), 520–527, Springer, Cham (2014).

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 3rd International Conference on Learning Representations – ICLR, (San Diego, CA, USA) (2015).

[30] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," BMC Medical Imaging **15** (Aug. 2015).

[31] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, **LNCS 9901**, (Cham, Switzerland), 424–432, Springer International Publishing (Oct 2016).

[32] M. L. McHugh, "Interrater reliability: the kappa statistic," Biochemia Medica **22**, 276–282 (Oct. 2012).

TABLE I: Confusion matrix between consensus labels assigned to baseline (level-sets) and predicted (U-nets) masks for the dataset S_1. In gray, CT scans for which the annotation remained unchanged. In green, images for which predicted masks received a better label. In red, images for which baseline masks received a better label.

Majority vote \ Level-sets	FS	mE	ME	FF	TOTAL
FS	6	7	4	2	19 (21%)
mE	7	20	16	11	54 (58%)
ME	1	8	6	4	19 (21%)
FF	0	0	0	0	0 (0%)
TOTAL	14 (15%)	35 (38%)	26 (28%)	17 (19%)	92 (100%)

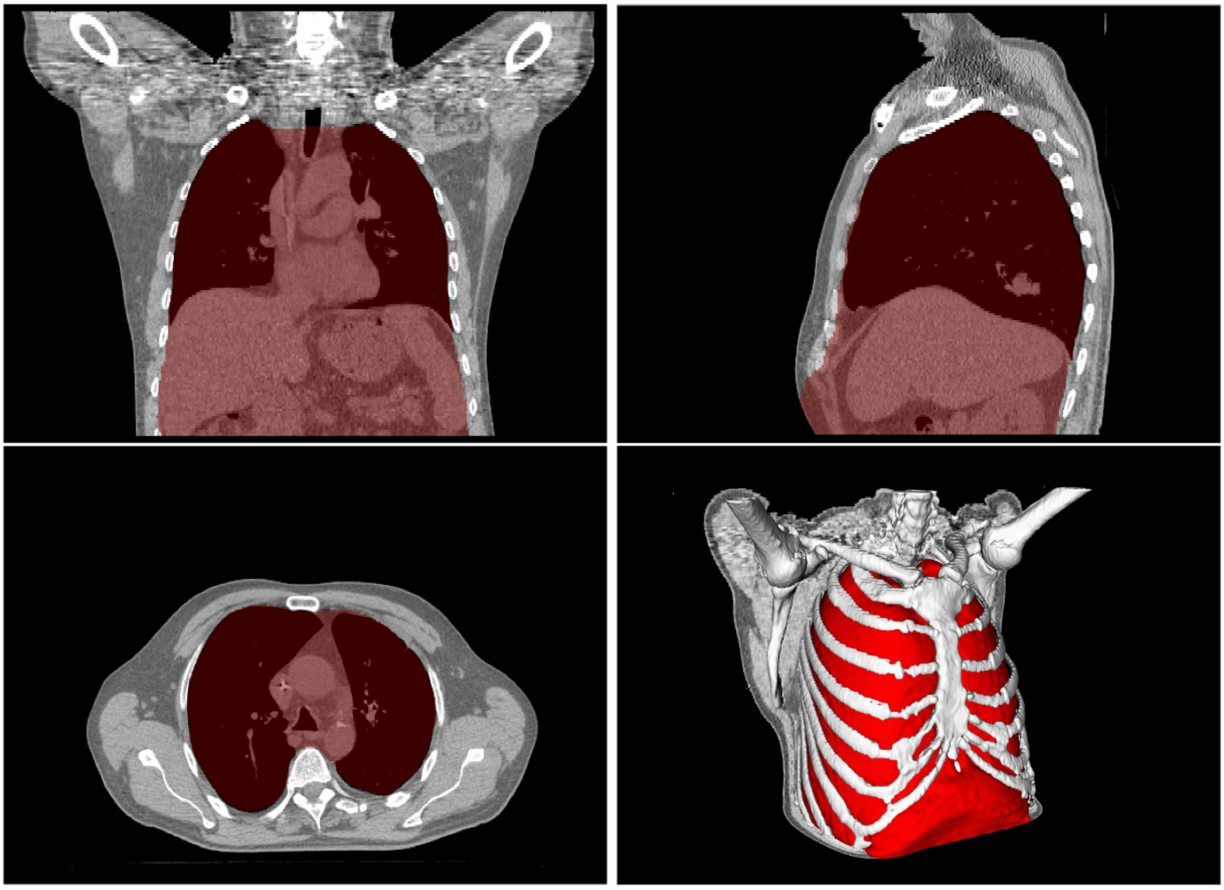

FIG. 1: Example of motion mask represented in 3D (in red color), as well as in translucent overlay on orthogonal views of a thoracic scan.

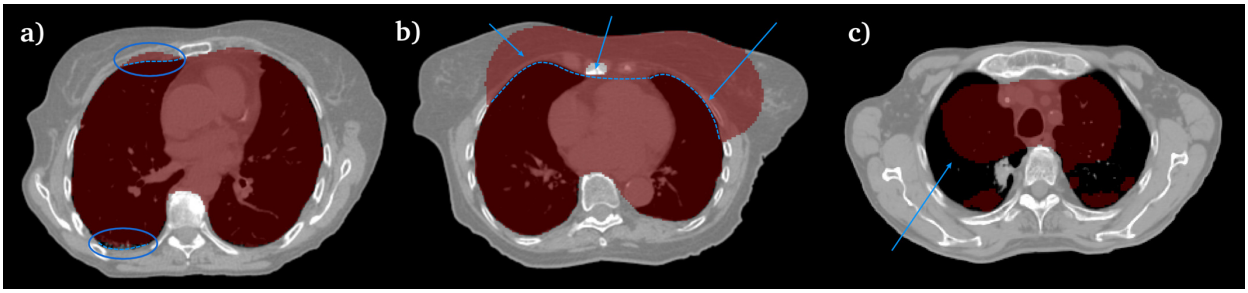

FIG. 2: Example errors found in segmented motion masks. Depending on their extent and location, these were labeled as minor error (a), major error (b), or full failure (c)

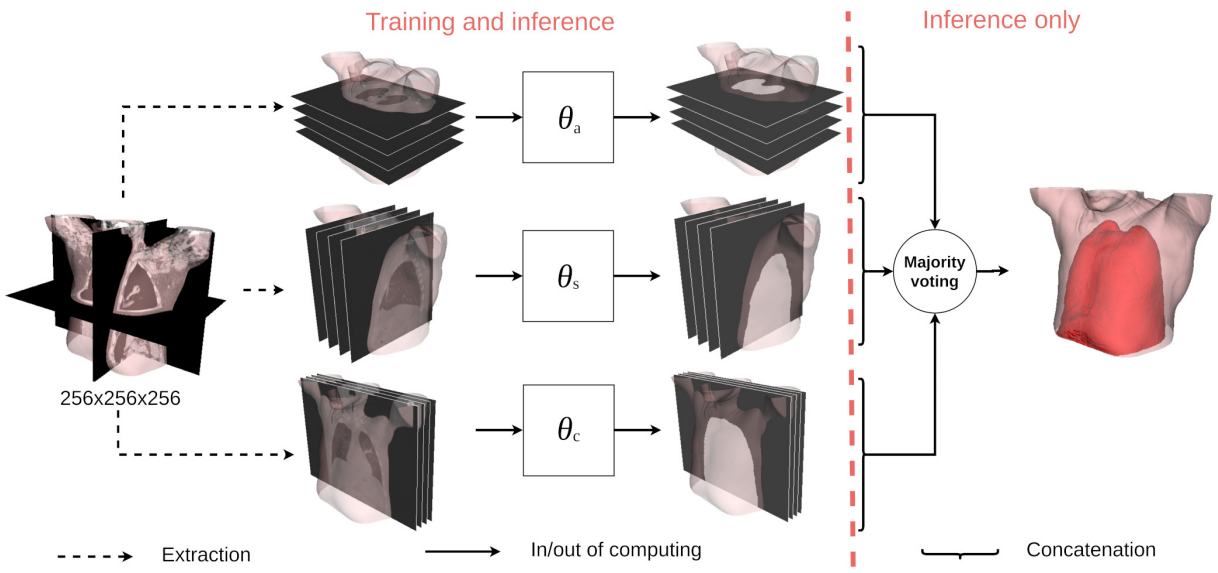

FIG. 3: In the proposed network architecture, the three U-nets θ_a, θ_c and θ_s are trained with 2D slices extracted from a volumetric image. During the inference, the 2D slices are concatenated to obtain 3D volumes and then merged using majority voting.

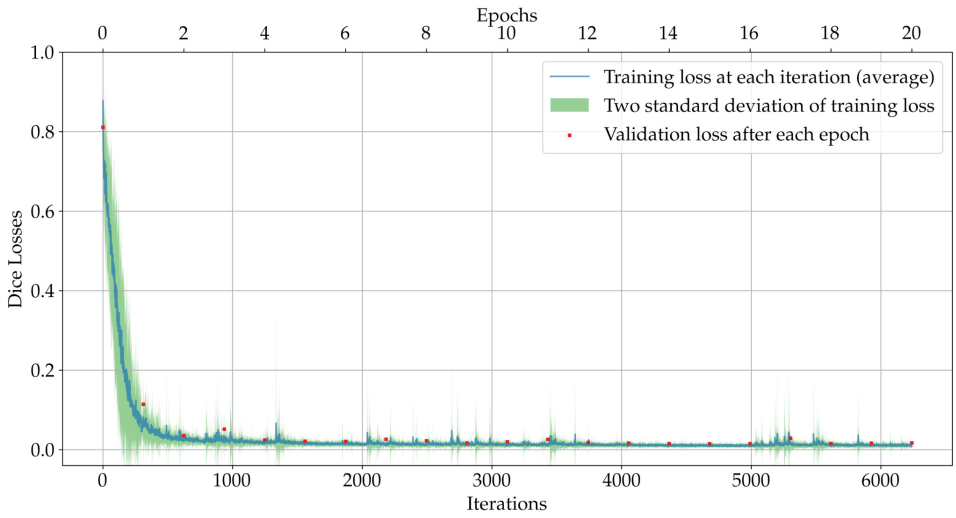

FIG. 4: Training and validation losses for the sagittal U-net (θ_s). The former are represented as mean value $\pm$ two standard deviations over the folds

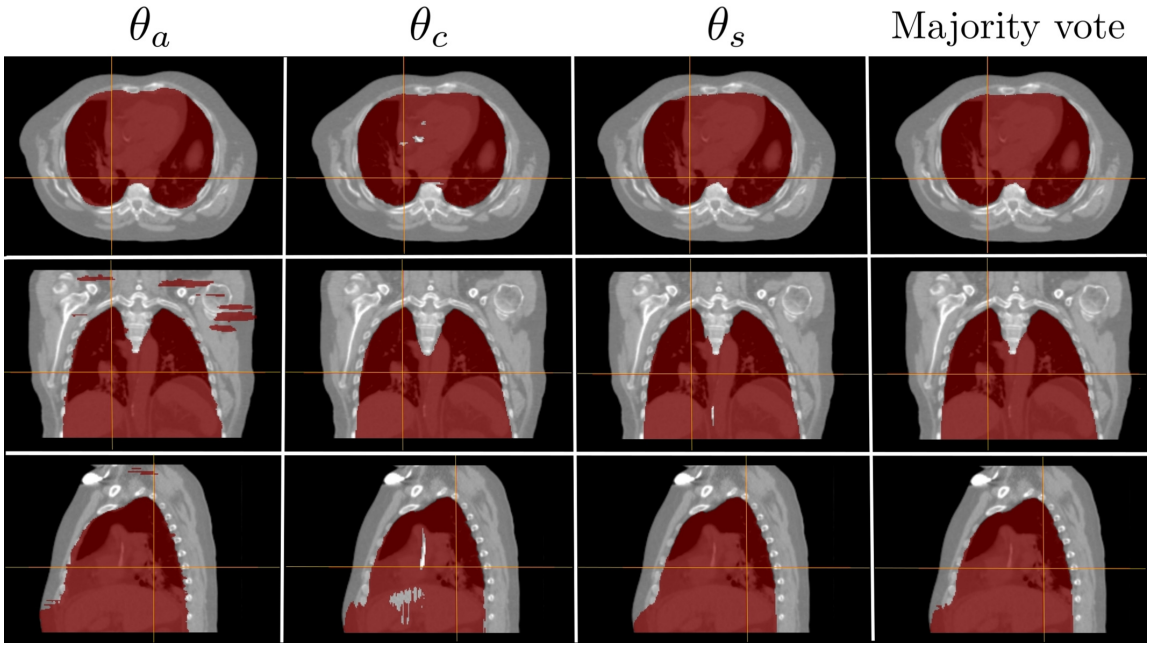

FIG. 5: Comparison of motion masks predicted by 2D U-nets trained on three types of orthogonal slices (from left to right: axial, coronal, and sagittal) and by majority vote (right-most), superimposed onto respective slices from one patient: axial (top), coronal (middle), and sagittal (bottom). Cross-hairs indicate the locations of the remaining slices.

TABLE II: Quantitative comparison between the full-success motion masks segmented by the baseline method (consensus label FS, $n = 14$), and those predicted by the 2D U-nets. Each measure is reported as mean value ± standard deviation, and the best result in each row is highlighted by a bold font. The last column reports the mean improvement with respect to the best 2D U-net.

measure	axial U-net θ_a	coronal U-net θ_c	sagittal U-net θ_s	majority vote	improvement
Dice (%)	98.50 ± 0.62	98.46 ± 0.49	98.70 ± 0.44	$\mathbf{98.96 \pm 0.32}$	0.3%
ASSD (mm)	2.1 ± 2.4	1.9 ± 1.7	1.1 ± 0.4	$\mathbf{0.8 \pm 0.2}$	23.0%
Hausdorff (mm)	87.7 ± 64.5	104.8 ± 72.2	61.9 ± 35.7	$\mathbf{37.6 \pm 26.7}$	39.2%

TABLE III: Confusion matrix between consensus labels assigned to baseline (level-sets) and predicted (U-nets) masks for the dataset S_2. In gray, CT scans for which the annotation remained unchanged. In green, images for which predicted masks received a better label. In red, images for which baseline masks received a better label.

Majority vote \ Level-sets	FS	mE	ME	FF	TOTAL
FS	2	9	3	0	14 (18%)
mE	2	25	21	1	49 (65%)
ME	0	2	11	0	13 (17%)
FF	0	0	0	0	0 (0%)
TOTAL	4 (5%)	36 (48%)	35 (46%)	1 (1%)	76 (100%)