Fusion of multiple segmentations of medical images using OV^{2}ASSION and Deep Learning methods: Application to CT-Scans for tumoral kidney

Lisa Corbat[a,*], Julien Henriet[a], Yann Chaussy[b], Jean-Christophe Lapayre[a]

[a] *FEMTO-ST Institute, DISC, CNRS, Univ. Bourgogne-Franche-Comté, 16 route de Gray, 25030 Besançon, France*
[b] *Centre Hospitalier Régional Universitaire Jean Minjoz, 3 boulevard Fleming, 25030 Besançon, France*

Abstract

Nephroblastoma is the most common kidney tumour in children. Its diagnosis is based on imagery. In the SAIAD project, we have designed a platform for optimizing the segmentation of deformed kidney and tumour with a small dataset, using Artificial Intelligence methods. These patient's structures segmented by separate tools and processes must then be fused in order to obtain a unique numerical 3D representation. However, when aggregating these structures into a final segmentation, conflicting pixels may appear. These conflicts can be solved by IA techniques. This paper presents a synthesis of our segmentation contribution in the SAIAD project and a new fusion method. The segmentation method uses the FCN-8s network with the OV^{2}ASSION training method, which allows segmentation by patient and overcomes the limited dataset. This new fusion method combines the segmentations of the previously performed structures, using a simple and efficient network combined with the OV^{2}ASSION training method as well, in order to manage eventual conflicting pixels. These segmentation and fusion methods were evaluated on pathological kidney and tumour structures of 14 patients affected by nephroblastoma, included in the

[*] Corresponding author, email address: lisa.corbat@univ-fcomte.fr
Email addresses: lisa.corbat@univ-fcomte.fr (Lisa Corbat), julien.henriet@univ-fcomte.fr (Julien Henriet), ychaussy@chu-besancon.fr (Yann Chaussy), jean-christophe.lapayre@univ-fcomte.fr (Jean-Christophe Lapayre)

final dataset of the SAIAD project. They are compared with other methods adapted from the literature. The results demonstrate the effectiveness of our training method coupled with the FCN-8s network in the segmentation process with more patients, and in the case of the fusion process, its effectiveness coupled with a common network, in resolving the conflicting pixels and its ability to improve the resulting segmentations.

Keywords: Medical image segmentation, Medical image fusion, Deep Learning, Conflict management, Cancer tumour

1. Introduction

The Wilms tumour, also called Nephroblastoma, is one of the most frequent abdominal tumours observed in children (generally from one to five years of age) and represents 5% to 14% of malignant paediatric tumours. This type of tumour is developed from embryonic structures in the kidney. Most often, its initial diagnosis is based on imagery. Generally, ultrasound observations are initially planned in order to confirm the tumour's existence and to approximate its position. A CT scan (Computed Tomography scan) then locates it, along with the affected organs and healthy tissues, with greater accuracy. Radiologists and surgeons need 3-Dimensional (3D) representations of the tumour and the border organs in order to confirm the diagnosis, plan the surgery (estimated quantity of blood loss, specialized equipment required, estimation of the duration of the surgery, anticipation of surgical risks, etc.) and also guide the actions of the surgeon during the surgery. This 3D representation is currently done through manual or semi-automatic segmentation, which is a long and time-consuming task.

The French-Swiss border project SAIAD (Automated Segmentation of Medical Images Using Distributed Artificial Intelligence) aims at obtaining automatic segmentation of the nephroblastoma and other abdominal structures through Artificial Intelligence (AI) methods. However, nephroblastoma is a rare tumour (the number of patients being limited, with for example the acquisition of data from 14 patients in 14 years at the University Hospital of Besançon), which can reach patients of various ages (between 1 and 15 years old). The structures, which already vary in size depending on age, also vary between patients. Thus, the size of the tumour, which is principally rounded in shape, can vary greatly from one patient to another, and the kidney affected, compressed by the tumour, also has a very varied shape and size. Solutions must be found to overcome the lack of data and the wide variety of structures encountered.

AI is a powerful tool that may provide a viable solution for fully automated treatment, but with all the constraints of nephroblastoma segmentation, a single AI method is not efficient enough to achieve correct segmentation for all abdominal structures. This is why each structure of the SAIAD project is segmented separately by a technique specifically adapted to a structure, and adapted to the limited data [1].

Once the segmentation step is completed for each structure, the different segmentations must be aggregated together to obtain a final consensus segmentation. This aggregation is not easy to carry out, because when the different segmentations are superposed, some areas of conflicting segmentations can appear on the labelled pixels belonging to the different structures. Other classical fusion techniques aggregated in a model have already been tested within the framework of the project [2].

1.1. Semantic segmentation

Many methods exist for image segmentation through Deep Learning and more particularly through CNNs, whose notions are presented in [3]. The first Fully Convolutional Network (FCN) was designed by Long et al. [4], where the fully connected layers are replaced by convolutional layers. SegNet [5] is a convolutional network that was developed to perform segmentation of indoor scenes and road scenes in real time. DeconvNet (Deconvolution Network) [6] is a CNN whose principle consists of a convolutional network followed by a deconvolutional network. DeepLab [7], revisits atrous convolution in order to solve the problem of segmenting objects at multiple scales.

In medical and biology applications, Thong et al. [8] used CNN to perform segmentation of healthy kidneys. U-Net [9] performed segmentation of cells in microscopy images. Currently, CNNs obtain accurate results on the recognition of the shape of a healthy kidney, because the shapes and areas are more or less the same from one subject to another. But there are gaps for the segmentation of more complex structures with very different forms depending on the patient in the literature. Neural networks for segmentation also need a large number of heterogeneous data in order to be able to transcribe reliable results. However, nephroblastoma is not a common tumour and the number of data is limited. In addition, pathological kidneys have very different forms from one patient to another, with unpredictable shapes and situations. This then us to find another method to segment more complicated structures with limited data.

In our previous contribution [1], the nephroblastoma was segmented by a CNN (FCN-8s) associated with the OV^{2}ASSION training method (called the FCN-8s-OV^{2}ASSION segmentation method), specially designed for training on a small dataset, and tested on a single patient. The pathological kidney was segmented by a Cased-Based Reasoning (CBR) coupled with a region growing technique, tested on a few slices of one patient.

1.2. Multiple segmentation fusion

The combination of multiple segmentations is less frequent than the segmentation process in the literature. Nevertheless, segmentation fusion (several segmentations of the same source image, calculated by more basic segmentation algorithms) can lead to better segmentations, instead of using a unique and complex segmentation method. It can also be used in the case of fusion of different complementary structures, as in our case, in order to arbitrate the conflicting pixels and obtain the final segmentation containing all the structures. Intuitive methods can be employed, such as the use of majority vote [10] or the intersection and the union [11]. However, these methods are limited. Many methods have emerged with the use of different metrics, via an iterative algorithm (Iterated conditional modes) with the Variation of Information

(VoI) criterion [12, 13]; with the F-Measure (or precision-recall criterion) [14]; with the Global Consistency Error (GCE) [15]; or the probabilistic Rand Index (PRI) measure, for the fusion of multiples segmentations [16]. Another approach is the use of spatial and intensity information (like the pixel's grey level and the neighbour's labels) of an image and its segmentations for the fusion of MR-T2 brain images segmentation map [17]. Recently, a new fusion method based on different models, which combined different fusion methods, has been used for the fusion of the nephroblastoma and the pathological kidney structures and is tested on three patients [2]. These different methods use different metrics (like the VoI and Dice metrics) and different information present in the CT-scans and segmentations such as pixel intensity and Euclidean distances between pixels.

The fusion of multiple segmentations using neural networks is a little-explored area of research. Most methods perform the fusion at the same time as the segmentation process in the same network. This is the case of Hu et al. [18], which uses a deep convolutional neural network consisting of seven convolutional layers, five ReLU layers, and two pooling layers that are used in order to combine two feature maps in an image segmentation method.

In addition, most methods apply fusion in the context of saliency detection. Saliency detection aims to highlight and segment the most important or visually distinctive objects or regions in an image by extracting its discriminative features and then computing their importance in the image. Li et al. [19] proposed a saliency detection approach that combines the segmentations from a pixel-level saliency map and a region-level saliency map using a single convolutional layer with a 1×1 kernel. The same fusion system is also proposed to fuse segmentations from five saliency maps [20]. In the saliency detection method proposed by Tang et al. [21], the fusion of two segmentations, each from a pixel-level saliency map and a region-level saliency map, is also used. The authors also incorporate the original image in the fusion process in order to provide more information and correct the segmentation in the final saliency map. The fusion system's architecture consists of one concatenation layer and three convolutional layers, outside of the segmentation process. Xiao et al. [22] developed another saliency detection method that uses a Recurrent Convolutional Neural Network to extract four saliency maps from an image. These maps are later concatenated and fused by three convolutional layers, followed by a Rectified Linear Unit (ReLU) layer. Another saliency method [23] fuses multiple saliency feature maps through a more complex convolutional neural network (outside of the segmentation process), and the resulting segmentation is passed through a Laplacian propagation to enforce better spatial consistency in the final saliency map. But these fusion networks are not applied at the fusion of complementary segmentations.

All these methods are present in the comparative Table 1 with more details. To simplify the networks architecture used, we use $conv(N, K)$, $pool(T, K)$, $deconv(N, K)$, $fc(N)$, sig, $relu$ and $dropout$ to indicate the different layers with N the number of output, K the kernel size and T the type of pooling layer.

Table 1: Comparison of the different fusion methods present in literature.

Classical fusion methods	**Working level**	**Calculation specificity**
by VoI metric [12]	pixel-level	Classical entropy + Mutual information
by F-Measure metric [14]	region-level	Precision measure + Recall measure
by GCE metric [15]	region-level	Local refinement error (LRE)
by PRI metric [16]	pixel-pair-level	Rand Index (RI)
by Feng et al. [17]	pixel-level	Euclidean distance + Greyscale intensity
based on multiple methods [2]	pixel-level	Aggregation of multiple fusion methods + Fusion of results by majority vote

IA fusion methods	**Application field**	**Network**
by Hu et al. [18]	Interactive segmentation	$conv1(512, 3) - relu1 - conv2(512, 3) - relu2 - conv3(512, 3)$ $-relu3 - pool3(MAX, 2) - conv4(4096, 7) - relu4$ $-conv5(4096, 1) - relu5 - conv6(1, 1) - deconv(1, 64)$
by Li et al. [19]	Saliency detection	$conv1(1, 1)$
by Tang et al. [21]	Saliency detection	$conv1(64, 3) - conv2(128, 3) - conv3(1, 1)$
by Xiao et al. [22]	Saliency detection	$conv1(32, 3) - relu1 - conv2(64, 3)$ $-relu2 - conv3(1, 3) - relu3$
by Qu et al. [23]	Saliency detection	$conv1(6, 5) - sig1 - pool1(MEAN, 2) - conv2(12, 5)$ $-sig2 - pool2(MEAN, 2) - conv3(24, 3) - sig3$ $-fc4(200) - relu4 - dropout4 - fc5(2)$

This paper proposes a synthesis of our segmentation contributions and a new fusion system based on neural networks, tested on the final set of 14 patients at our disposal in the SAIAD project:

- We show the performance of our neural network segmentation system, designed to operate per patient on a small dataset using our training method called OV^{2}ASSION. This time, we are testing this segmentation system on the structure of the tumour and the pathological kidney of our patients;
- We then propose a new neural network fusion method, by designing a simple efficient network, inspired by classical segmentation networks, composed of an upsampling layer;
- We also use the OV^{2}ASSION specific training method in order to test the fusion of the tumour and pathological kidney segmentations of each patient, with a small training dataset;
- Segmentation and fusion systems are tested for the first time on 14 patients (previously only one patient had been tested with the segmentation system).

This paper is organized as follows: Section 2 presents the general architecture of the SAIAD project, with a description of the segmentation method and the new fusion method; Section 3 presents the experiments and the performances of the segmentation and fusion methods; lastly, these results are finally discussed in Section 4.

2. Proposed method

This part of the paper first presents the general architecture of the platform of the SAIAD project, and more particularly its segmentation and fusion layers. The way of the OV^{2}ASSION method, used in the segmentation method, has been integrated into the fusion of multiple segmentations is then presented.

2.1. The SAIAD project's architecture

Figure 1 shows the general architecture of the system designed in the SAIAD project. It is composed of three layers. The first one is the data layer, which contains a database for each segmentation system. Each database has access to all of the CT scan images (Atlas) and expert knowledge, such as the corresponding manual segmentations (ground truths) and the data and metadata of CT scans stored as Dicom files. The second layer is the segmentation layer, where CT scans are segmented by AI systems. The nephroblastoma and kidney segmentations can be performed independently by a Deep Learning (FCN-8s) coupled with the OV^{2}ASSION method [1]. The CBR coupled with a region growing system, also used for the pathological kidney segmentation in [1], does not currently allow the segmentation of an entire patient. Subsequently, this method is therefore not used to obtain the segmentations of a patient's pathological kidney (before being fused with the segmentations of his tumour). A Conditional Random Field (CRF) post-processing [24] is applied, allowing improvement of the segmentations. At the end of the segmentation processes, the system gives two complementary segmentations in two different images and the fusion layer combines them with a deep neural network.

In addition to the work done in [2] (The creation of new models for the fusion of our two structures, which use a combination of different fusion methods), we have now used a Deep Learning fusion method, with the OV^{2}ASSION training method to train the neural network used in this last fusion step of the process.

2.2. Overlearning Vector for Valid Sparse SegmentatIONs (OV^{2}ASSION) method for segmentation and the fusion

Having a sufficiently large number of data representative of all possible data is essential for training a deep neural network. Manual segmentation of one patient is a time-consuming process (six to eight hours for an expert). Consequently, at the scale of a hospital, the learning set composed of the entire segmented abdomens of patients may be composed of tens of cases only. This

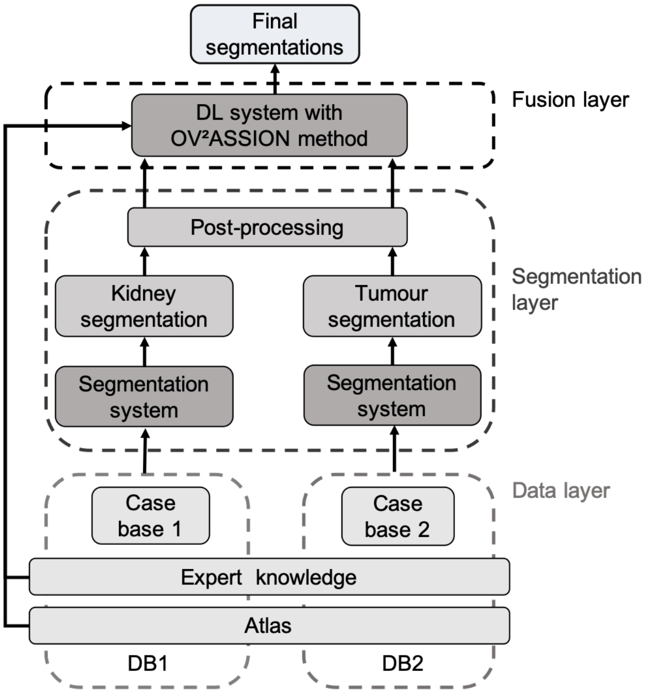

Figure 1: General schema of the SAIAD project.

may not be large enough for conventional learning since each tumour and pathological kidney is unique and varies greatly from one patient to another. Nevertheless, for any given patient, these structures have the same shades of grey and are generally homogeneous. As shown in Figure 2, the OV^{2}ASSION method is based on the overlearning of some manually segmented slices of the patient, separated by a gap in order to calculate the segmentation of the entire set of unsegmented slices of this patient automatically.

In this particular training method, we need to determine the training dataset and the testing dataset. This datasets are formalized in the format of a vector noted V_g which determines for each slice of a patient, the slices taken into account for training and those for testing, all according to the gap g. The possible values for a slice are 1 if it is taken into account for training and 0 if it is taken into account for testing.

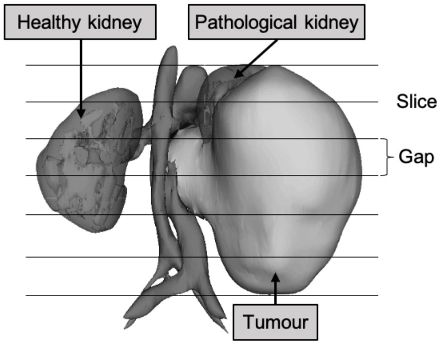

Figure 2: The 3D representation of the tumour and pathological kidney. Each black line represents the selected slice for the training of the neural network. The gap between the chosen slices is the same to recover information homogeneously at different levels.

For a given gap, there are several possibilities for the set of selected slices for training, depending on the first slice considered for training. Consider a V_g vector with a g gap and being part of V (corresponding to the set of all possible combinations), such that:

$$V_g \in V = \{(\{0,1\}_1, \ldots, \{0,1\}_n)\} \tag{1}$$

Where n is the number of the patient's slices.

The learning set LS of the neural network, whether it is a network for segmentation or fusion, then contains all the slices used for training that have been determined by a vector, and is defined as:

$$LS = \{S_j, \ldots, S_k\} \tag{2}$$

Where $S_j, \ldots, S_k$ represent all the selected slices for training.

However, there are different possible vectors for a given gap, depending on the first slice included in LS. V_g then corresponds to the set of possible vectors as a function of g, such as:

$$V_g = \bigcup_{i=1}^{h} (V_g)_i \tag{3}$$

Where h is the number of possible vectors for the g gap and $(V_g)_i$ is the i number vector of the g gap.

Thus, the $(V_g)_1$ vector corresponds to a vector starting by including the first slice at LS before including all $g+1$ slices, the $(V_g)_2$ vector starting by including the second slice at LS, and so on. An example of these different vectors on 60 slices for a 2 gap is shown in Table 2 where three vectors are then possible.

Table 2: Example of possible vectors for a gap of 2 $(V_2)_i$ and for 60 slices of a patient ($n = 60$).

	S_1	S_2	S_3	S_4	S_5	...	S_{58}	S_{59}	S_{60}
$(V_2)_1$	1	0	0	1	0	...	1	0	0
$(V_2)_2$	0	1	0	0	1	...	0	1	0
$(V_2)_3$	0	0	1	0	0	...	0	0	1

The formalization of the addition of slices in LS can be done by a f function which allows to select a slice having a value of 1 in a vector $(V_g)_i$:

$$f \ : \ LS \times \{0,1\} \rightarrow LS | \forall i \in \{1,\ldots,n\} \begin{cases} f(S_i, 0) = \emptyset \\ f(S_i, 1) = S_i \end{cases} \tag{4}$$

And finally we can define the final LS, by introducing a F function that uses the f function following the $(V_g)_i$ vector, going through all the elements of the vector.

$$F \ : \ V \rightarrow LS | F(V_g)_i) = \bigcup_{j=1}^{n} \{f(S_j, ((V_g)_i)_{j-1})\} \tag{5}$$

With this method, the trained dataset is composed of selected slices with a certain gap between them, and the test dataset is composed of the rest of the patient's slices.

In our case, the different structures have only a slight difference from one neighbouring slice to another. This semi-automatic method can be effective for the fusion of multiple segmentations. Nevertheless, the OV^{2}ASSION implementation for this fusion step will consider the manual segmentations, which have already been considered during the segmentation step before fusion.

2.3. The DL network for the fusion

For the fusion of multiple complementary segmentations with the method indicated above, a common neural network is used. This neural network, whose general architecture is presented in Figure 3, is inspired by the CNN used for image segmentation and more particularly by the networks of Long et al. [4]. The network for the fusion is composed of eight convolutional layers and two pooling layers. It consists of the application of two 3x3 convolutions, each followed by a rectified linear unit (ReLU), and a 2x2 max pooling with a stride of 2 for downsampling, with the whole process repeated twice. Then, two other 3x3 convolutions without padding with ReLU are applied and followed by a 1x1 convolution in order to obtain the desired number of classes. Finally, a 20x20 up-convolution (or deconvolutional layer) is used to get a result with the same size as the input of the network. For the training phase, like the training for segmentation, all chosen slices are used for the prediction of the segmentation fusion of the patient's entire nephroblastoma.

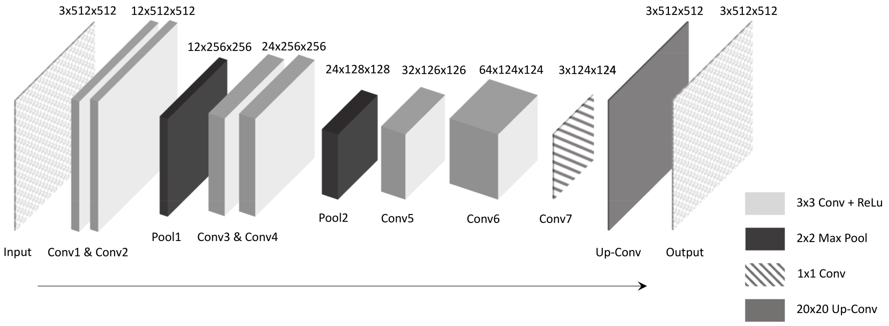

Figure 3: Overall architecture of the proposed network.

3. Experiments

For all the methods (segmentation and fusion) which use the OV^{2}ASSION training method, one training was performed per patient and for each structure. Likewise, a gap of 4 is used to select training slices.

The FCN-8s-OV^{2}ASSION segmentation method will be compared with the U-Net architectures with the OV^{2}ASSION training and the FCN-8s with a classical training (i.e. by performing training on all the other patients in order to allow segmentation of the current patient).

After, the pathological kidney segmentation and the tumour segmentation was fused by the new fusion system with the OV^{2}ASSION method, called the DL-OV^{2}ASSION fusion method, and compared by the fusion systems of Corbat et al. [2] and others of the literature.

All neural networks were developed in Python with the Caffe library [25], on a Tesla Kepler K40 GPUs from Mésocentre at the University of Franche-Comté.

3.1. Dataset and ground truth

CT-Scan of 14 nephroblastoma patients are used in the SAIAD project for the segmentation process and the segmentations fusion process. These patients come from the *Centre Hospitalier Régional Universitaire de Besançon* (University Hospital of Besançon) between 2005 and 2018. There are as many girls as boys and their age varies between 1 and 15 years, with an average age of 5 and a half years. The tumoral kidney can be either the left or the right kidney and the number of slices to visualize all the target structures vary from 30 to 147 per patient, with an average of 97 slices per patient (in total, we have 1344 slices for all patients).

Per patient, the training dataset and the testing dataset are the same in the segmentation stage and the fusion stage. Using the OV^{2}ASSION training method with a gap of 4 (starting with the first slice used for training) and an average of 97 slices per patient, we have an average of 20 slices for the training set and 77 slices for the testing set. That's about 20% of a patient's slices that are used for training network.

We had the ground truth segmentations of all the patients (i.e. the desired segmentations), carried out by experts (surgeons and radiologists) at the same hospital. All of the CT scan images and ground truths had the same size: 512 x 512 pixels. These ground truths were used in order to verify the reliability of our processes and the resulting segmentations through the Dice coefficient. The Dice's score varies between 0 and 1. Between two segmentations S and S', S represents the pixels of the ground truth, and S' represents the pixels of the calculated segmentation given by the system. A score of 0 denotes that the two segmentations are completely different, whereas a score of 1 indicates that they are identical.

The Dice is then defined as:

$$Dice(S, S') = \frac{2 * TP_{S,S'}}{2 * TP_{S,S'} + FP_{S,S'} + FN_{S,S'}} \tag{6}$$

where $TP_{S,S'}$ is the number of true-positive pixels between S and S'; $FP_{S,S'}$ is the number of false-positive pixels; and $FN_{S,S'}$ is the number of false-negative pixels.

In addition, the IoU metric (also known as the Jaccard index and the Hausdorff distance) is calculated in order to compare the results according to different metrics.

3.2. Tumour and pathological kidney segmentation

3.2.1. Training optimization and post-processing

For the FCN-8s-OV^{2}ASSION method, we used the pre-trained parameters of the conv1 to fc7 layer of the PASCAL VOC 2012 database [26]. The last convolution was modified with two-channel dimensions to predict two scores: the background and the chosen structure. We used a learning rate of 1e-12 and 10,000 iterations for the training. For other comparative methods, we used a learning rate of 1e-8 with 100,000 iterations for the training of traditional FCN-8s, and the same pre-trained parameters as FCN-8s-OV^{2}ASSION. We used a learning rate of 1e-9 with 10,000 iterations for the U-Net-OV^{2}ASSION, but the network was pre-trained on one patient during 75,000 iterations. A CRF post-processing [24] was applied at the end of each neural network for segmentation process.

3.2.2. Results of the segmentations

Initially, as the structure of the tumour is simpler to segment than the structure of the pathological kidney, we compared the effectiveness of our FCN-8s-OV^{2}ASSION method on the segmentation of the tumour with other methods. In particular, with FCN-8s architecture and traditional training, and with U-Net architecture coupled with OV^{2}ASSION training. Using the U-Net architecture with traditional training does not give any results because the network is deeper and there is not enough data to allow the network to train correctly.

Table 3 presents the mean results, in the different metrics, on 14 patients of tumour segmentation, by the 3 methods. We can see that our method on average achieves the best tumour segmentation, with a Dice and a IoU of 91.54% and 84.68% . The Hausdorff distance is also smaller with our training method coupled to the different networks. Traditional training with the FCN-8s network has not been conclusive, because in the absence of more consistent data, the network segmented some patients poorly, resulting in a high standard deviation. The U-Net-OV^{2}ASSION method also performs correct segmentation. However, the computation time for segmenting a patient is three times longer with this method than with the FCN-8s-OV^{2}ASSION.

Table 3: Mean results (in Dice score, IoU metric and Hausdorff distance between calculated segmentations and ground truths) on 14 nephroblastoma patients of the tumour segmentation by our FCN-8s-OV^{2}ASSION method, compared with the traditional FCN-8s and the U-Net with the OV^{2}ASSION training method.

	Dice ± Std Dev	**IoU ± Std Dev**	**HD ± Std Dev**
FCN-8s-OV^{2}ASSION method	**0.9154** ± 0.0449	**0.8468** ± 0.0726	4.27 ± **1.41**
U-Net-OV^{2}ASSION method	0.8994 ± **0.0429**	0.8197 ± **0.0687**	**4.12** ± 1.43
FCN-8s	0.6708 ± 0.3448	0.5828 ± 0.3264	5.51 ± 2.13

Table 4 presents the Dice scores obtained of both structures for each patient, as well as the standard deviation for the FCN-8s-OV^{2}ASSION method. This Dice score per patient was obtained by calculating the average of the Dice scores of all of the patient's slices. We find again a Dice score of 91.54% on average for the tumour segmentation, with a standard deviation of 4.49%. The pathological kidney segmentation is correct at 80.58% with a higher standard deviation of 11.27%. On average, with both structures, the Dice score is 86.06% with 4.72% of standard deviation. We can see that the segmentation system is more efficient, on average, on tumour segmentation than on the kidney, with a difference of 11%. We have noticed that our system segments larger structures better. Thus, Patients 3 and 14, having a smaller tumour compared to all the other patients, have the lowest Dice score for the tumour at 80.66% and 84.13%, and Patient 10, having the kidney most compressed by the tumour and therefore less visible, has the lowest Dice score for the kidney at 61,81%. In addition, the tumour is often oval in shape and larger in size, allowing more accurate segmentation, whereas the kidney, completely compressed, will have more random and less regular shapes. This point can be seen in Figure 4, which shows the results of the segmentation of both structures for one slice of Patient 10.

Table 4: Results of the tumour and pathological kidney segmentation (in Dice score) through the FCN-8s-OV$\overset{2}{\mathring{A}}$SSION method.

Patient	Tumour segmentation	Pathological kidney segmentation	Mean
P1	0.9368	0.8553	0.8961
P2	0.8694	0.9115	0.8904
P3	0.8066	0.9346	0.8706
P4	0.9527	0.9187	0.9357
P5	0.9290	0.8433	0.8862
P6	0.9202	0.8295	0.8748
P7	0.9500	0.7254	0.8377
P8	0.9203	0.9087	0.9145
P9	0.9452	0.6686	0.8069
P10	0.9282	0.6181	0.7731
P11	0.9588	0.7353	0.8471
P12	0.9386	0.6376	0.7881
P13	0.9183	0.7693	0.8438
P14	0.8413	0.9258	0.8835
Mean	0.9154	0.8058	0.8606
Std Dev (%)	4.49	11.27	4.72

3.3. Fusion of the tumour and pathological kidney segmentation

3.3.1. Training and prediction phase

The network (presented in Figure 3) needs to receive as inputs both of the segmentations to fuse (pathological kidney and nephroblastoma) and the corresponding CT scan in order to give as much information as possible. These images are contracted in a single 3-channel image, which is given as input to the network. When considering one patient, the network trains according to the OV$\overset{2}{\mathring{A}}$SSION method on all of the selected slices of the patient with a gap of

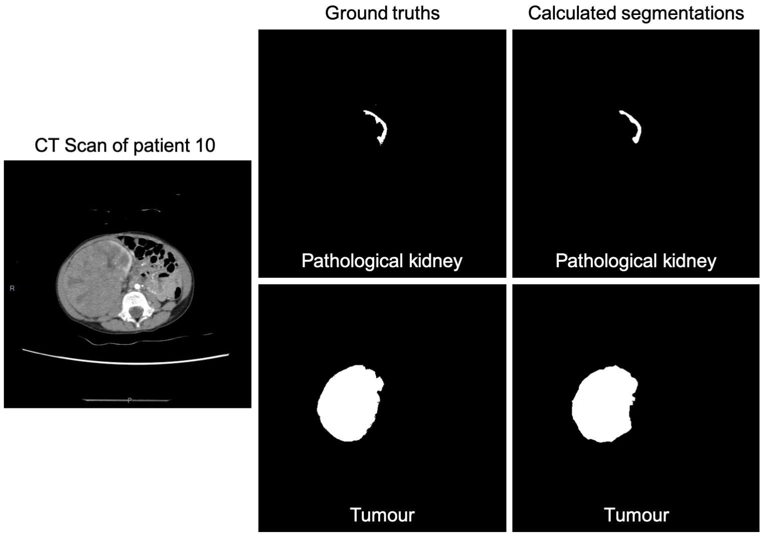

Figure 4: Segmentation result of one slice of Patient 10. From left to right: CT scan image of patient; Ground truths of the pathological kidney and the tumour; Result of the pathological kidney segmentation and tumour segmentation obtained by AI method.

4, starting with the first slice. Indeed, these parameters have been found to be the ones that gave the best compromise between accuracy and time consumed for manual segmentation [1]. The deep neural network for our DL-OV$\overset{2}{\text{Å}}$SSION fusion system is trained on 30,000 iterations with a learning rate of 1e-9. For the prediction phase, all of the slices of the considered patient (even the slices used for training) are given to the network.

3.3.2. Results of the segmentation fusion

We, therefore, tested the segmentation fusion, based on the segmentations previously acquired by the FCN-8s-OV$\overset{2}{\text{Å}}$SSION method, on all 14 patients. Our results are compared to seven other methods, including four neural network fusion methods (a network of three convolutional layers [21]; a network of three convolutional layers followed by ReLU layers [22]; and a deeper neural network [23]) and three so-called more classical methods (the use of a combination of six mathematical techniques and metrics [2]; the use of similarity criterion [17]; and the use of VoI-criterion [12]). All fusion methods based on neural networks (except the DL-OV^{2}ASSION) have been subjected to classical training (i.e. by performing training on all the other patients in order to allow fusion of the current patient) and trained with a learning rate of 1e-10 and the number of iterations fixed at 20,000.

Table 5 presents the percentage of conflicting pixels that have been correctly classified and the standard deviation of the different approaches. The DL-OV^2 ASSION method is not the most efficient in the management of conflicting pixels but is in second place for conflict resolution. However, this method performed the best conflict resolution for four patients. It is the method that uses a combination of six different mathematical methods that obtain, on average, the best pixel resolutions in conflict (with the best conflict resolution for nine patients). Nevertheless, when examining the standard deviations obtained, the DL-OV^2 ASSION method has a higher standard deviation than most of the methods. It then obtains more dispersed results, as is the case for P11, where the network had difficulty in resolving conflicts. We can assume that this is a lack of training for this patient, as the training parameters are fixed for all patients.

Table 5: Percentage of correctly resolved conflicting pixels for each patient with different methods. The values in bold correspond to the best values obtained. From left to right: The different patients; The results of our DL-OV$\overset{2}{\text{Å}}$SSION method; The results by a deeper neural network; The results by the network of three convolutional layers; The results by the network of three convolutional layers followed by ReLU layers; The result of the method with a combination of six fusion protocols; The results by the use of similarity; The results by the use of VoI-criterion.

Patient	DL-OV$\overset{2}{\text{Å}}$SSION method	Network by Qu et al. [23]	Network by Tang et al. [21]	Network by Xiao et al. [22]	Method by Corbat et al. [2]	Method by Feng et al. [17]	Method by Mignotte [12]
P1	0.7002	0.6400	0.6041	0.6162	**0.7296**	0.6062	0.6782
P2	**0.7655**	0.6926	0.6795	0.6831	0.6989	0.5834	0.6681
P3	**0.9336**	0.7012	0.7054	0.0249	0.7220	0.6556	0.5104
P4	**0.8075**	0.2650	0.7189	0.7915	0.7332	0.6549	0.6650
P5	0.6527	0.5855	0.5655	0.6003	**0.7213**	0.5907	0.7089
P6	0.5834	0.5692	0.5815	0.6087	**0.7486**	0.6201	0.6044
P7	0.6891	0.6658	0.6819	0.6143	**0.7138**	0.5426	0.5870
P8	**0.7840**	0.3396	0.3758	0.3461	0.4295	0.3868	0.2846
P9	0.6242	0.6287	0.6399	0.6553	**0.8087**	0.6528	0.5250
P10	0.6490	**0.6658**	0.6175	0.5321	0.6618	0.5463	0.5471
P11	0.3152	0.4586	0.4574	0.6378	**0.7911**	0.5303	0.7108
P12	0.5708	0.4683	0.5656	0.5622	**0.6376**	0.5746	0.4759
P13	0.5965	0.6255	0.6222	0.4777	**0.7282**	0.6122	0.5563
P14	0.6539	0.5226	0.5788	0.6482	**0.7224**	0.6114	0.5697
Mean	0.6661	0.5592	0.5996	0.5570	**0.7033**	0.5834	0.5780
Std Dev (%)	14.23	13.35	9.38	18.42	9.00	**6.98**	11.32

Table 6 shows the results obtained according to three different metrics (Dice score, IoU metric and Hausdorff distance) between calculated segmentation and ground truth) for the tumour and the pathological kidney structures, and the mean of both, before and after the fusion, according to the different methods. The results (Dice, IoU, HD) for one method were obtained by averaging the results of all 14 patients, and the result for one patient was obtained by averaging the results of each slice of the patient. The DL-OV^{2}ASSION method obtained, on average, the best scores, either for the tumour or the pathological kidney segmentation, with a Dice for both structures of 88.06% and 4.65% of standard deviation. It also obtained higher scores than the scores before the fusion. Thus, the tumour segmentations improved, from a Dice score of 91.54% to 92.66% with a decrease in the standard deviation. The same result was obtained for the pathological kidney segmentations, going from a Dice score of 80.58% on average to 83.45%, with a decrease in the standard deviation as well. Finally, we obtained a general improvement of the segmentations, going from a Dice score of 86.06% on average to 88.06%, with a regular decrease in the standard deviation. The same improvement is noticed by the other metrics.

The method by Feng et al. and by Mignotte obtain very similar results compared to the results before the fusion because these methods do not allow for higher results, as they only work on the conflicting pixels on the image and do not modify the other pixels. Neural network fusion methods modify the overall segmentation, and can, therefore, modify the labels of all pixels. Moreover, the number of conflicting pixels in the segmentations is, on average, very low (1.73%) compared to the total number of pixels in the image, which is why the metric results can only show a slight improvement in the best case with this type of method.

Table 6: Results of the quality of the tumour segmentation, the pathological kidney segmentation and both, using Dice score, IoU metric and Hausdorff distance with standard deviation, before the fusion and after the fusion according to different methods.

	Dice ± Std Dev			**IoU ± Std Dev**			**HD ± Std Dev**		
	Tumour	P. Kidney	Mean	Tumour	P. Kidney	Mean	Tumour	P. Kidney	Mean
DL-OV$\overset{2}{A}$SSION method	**0.9266** ± **0.0373**	**0.8345** ± 0.1048	**0.8806** ± 0.0465	**0.8653** ± **0.0622**	**0.7165** ± 0.1632	**0.7909** ± 0.0711	**4.11** ± 1.40	**3.36** ± 0.70	**3.73** ± 0.81
Network by Qu et al. [23]	0.9217 ± 0.0384	0.8256 ± **0.0987**	0.8737 ± 0.0416	0.8570 ± 0.0634	0.7139 ± **0.1406**	0.7854 ± 0.0578	4.27 ± **1.31**	3.44 ± 0.66	3.85 ± **0.74**
Network by Tang et al. [21]	0.9163 ± 0.0448	0.8108 ± 0.1101	0.8635 ± 0.0452	0.8482 ± 0.0727	0.6950 ± 0.1539	0.7716 ± 0.0611	4.29 ± 1.39	3.51 ± 0.65	3.90 ± 0.76
Network by Xiao et al. [22]	0.9129 ± 0.0466	0.8118 ± 0.1009	0.8624 ± **0.0410**	0.8428 ± 0.0751	0.6944 ± 0.1413	0.7686 ± **0.0558**	4.42 ±1.38	3.68 ± 0.75	4.05 ± 0.83
Method by Corbat et al. [2]	0.9157 ± 0.0452	0.8065 ± 0.1141	0.8611 ± 0.0481	0.8473 ± 0.0730	0.6895 ± 0.1564	0.7684 ± 0.0637	4.25 ± 1.39	3.47 ± **0.64**	3.86 ± 0.76
Method by Feng et al. [17]	0.9154 ± 0.0453	0.8056 ± 0.1143	0.8605 ± 0.0481	0.8469 ± 0.0732	0.6884 ± 0.1566	0.7676 ± 0.0636	4.25 ± 1.39	3.47 ± **0.64**	3.86 ± 0.76
Method by Mignotte [12]	0.9154 ± 0.0456	0.8039 ± 0.1168	0.8596 ± 0.0494	0.8468 ± 0.0735	0.6865 ± 0.1592	0.7666 ± 0.0651	4.26 ± 1.40	3.48 ± **0.64**	3.87 ± 0.77
Dice before fusion	0.9154 ± 0.0449	0.8058 ± 0.1127	0.8606 ± 0.0472	0.8468 ± 0.0726	0.6883 ± 0.1548	0.7676 ± 0.0626	4.27 ± 1.41	3.49 ± 0.65	3.88 ± 0.77

Figure 5 shows the results according to the different methods or networks. Three examples of conflict management are presented. For each segmentation, there is a zoom on the area of conflict surrounded. The figure is divided into two columns: on the left the ground truth at the top and the resulting segmentation with conflicts at the bottom; On the right the different results. On the top right we have the results by the DL-OV^{2}ASSION method, by the Qu et al. network [23], by the Tang et al. network [21], by the Xiao et al. network [22], and bottom right the results by the Corbat et al. method [2], by the Feng et al. method [17] and by the Mignotte method [12]. For each results, the Dice score of the tumour and the pathological kidney are present as well as the percentage of good conflict resolution.

In the case of our three examples, our method obtains on average a better improvement of the quality of the segmentations. The segmentations are also smoothed and filled. On the other hand, when managing conflicting pixels, the results are more heterogeneous. In the first example, non-AI-based methods are the most successful in fusion. his is also the case in the second example, but some of the AI-based methods also achieve this. In the third example, none of the methods achieve this, except for the DL-OV^{2}ASSION method with a good resolution percentage of 93.48%. On average, compared to all the slices of all our patients, we have already shown that our method achieves a good resolution, but this is the Corbat et al. method [2] that gets the best results (Table 5).

Figure 6 shows the results of the segmentation fusion with conflict management of the DL-OV^{2}ASSION method. It shows, in particular, the results obtained on ten slices belonging to different patients. All pixels in conflict are labelled as belonging either to the tumour, kidney, or background. In the end, the fusion method using DL-OV^{2}ASSION significantly improves the segmentations, thus adding effective post-processing to the segmentation.

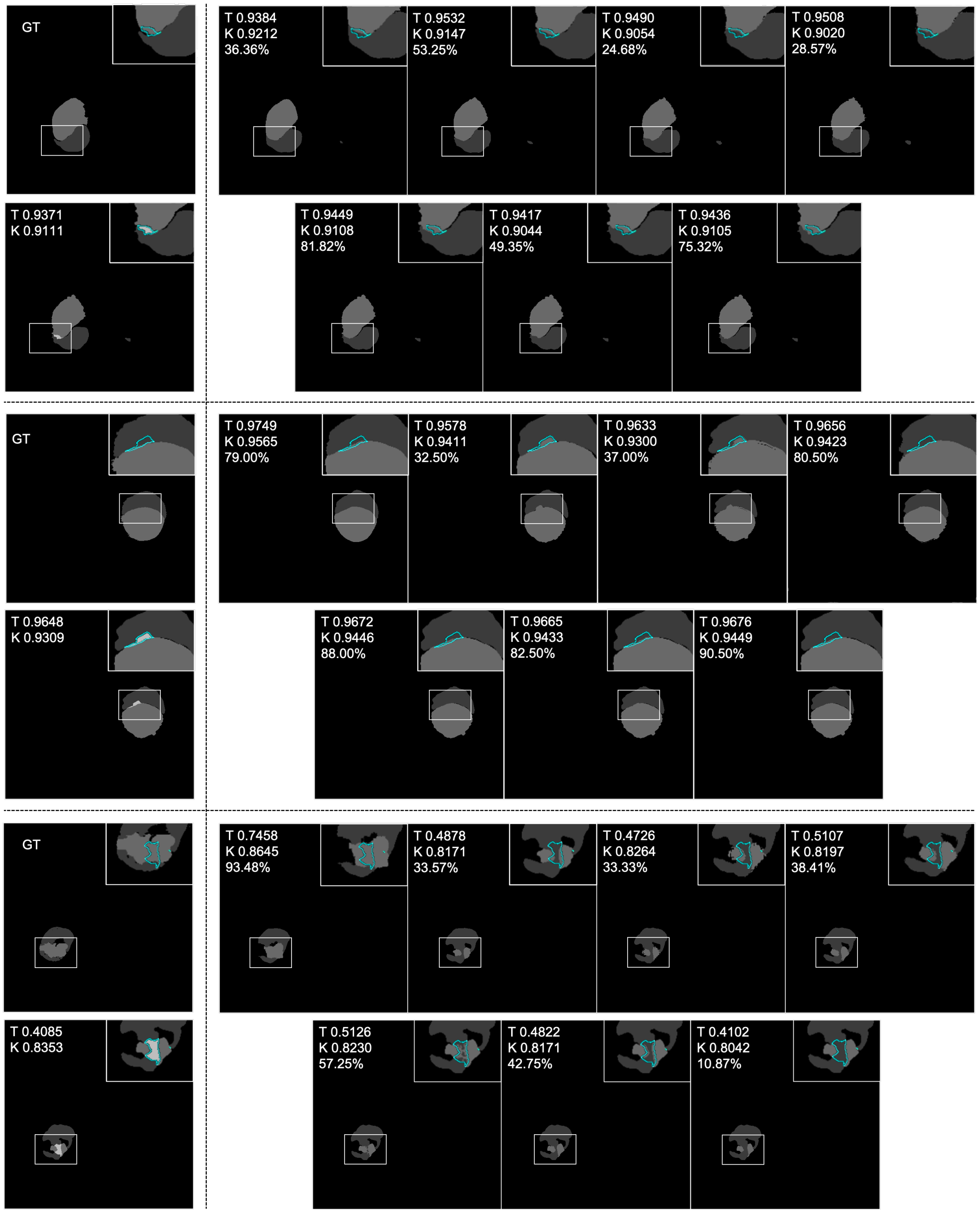

Figure 5: Comparison of different segmentation results and conflicting pixel management according to the network or method used. The elements from darkest to brightest represent: the background; the pathological kidney; the tumour; and the few conflicting pixels.

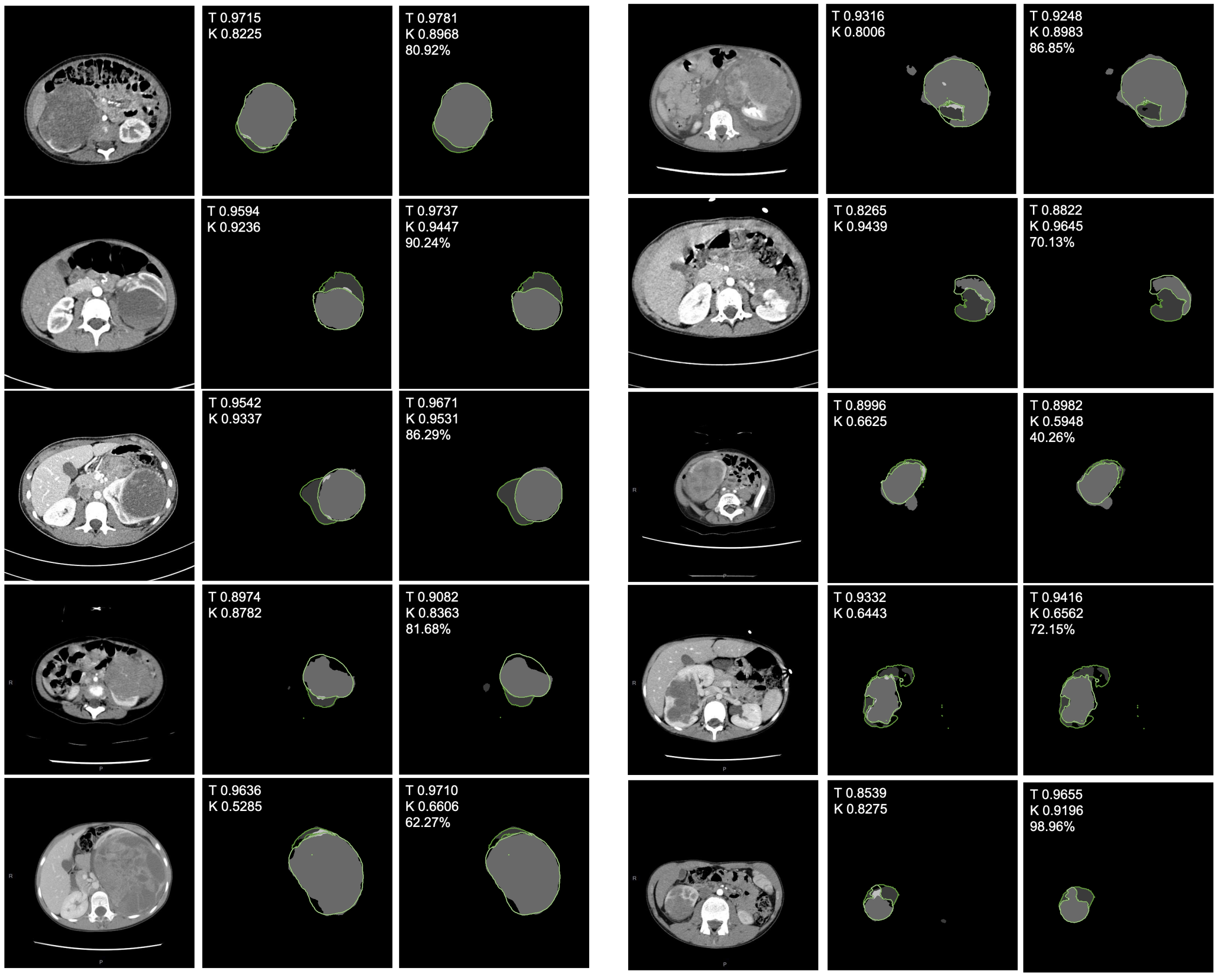

Figure 6: Results of the segmentation fusion of some slices by our DL-OV$\overset{2}{A}$SSION method. There are two columns of results. For each column, from left to right: CT scan images; Segmentations with conflicts; Results of the segmentations fusion. For the images representing the segmentations, the curves denotes the ground truth. The Dice score for each structure and the percentages of good conflict management are present for each segmentation.

4. Discussion

The presented DL-OV^{2}ASSION method achieves the highest percentage in terms of segmentation improvement, for the tumour structure as well as for the pathological kidney structure and regardless of the metrics used. But, the percentage of conflicting pixels that are correctly resolved, for each patient and each method, is not relatively high, because of the difficulty of the conflicts to be solved. Most conflict zones are located at the intersection of the different structures. These areas remain very ambiguous, and the delimitations between two structures are not always clear even for radiologists who then create the contours through their knowledge and experience.

In addition, this new method obtains a higher standard deviation compared to other methods. This is due to the difficulty of the neural network used to segment some patients correctly. For example, this is the case for Patient 11, who obtained a percentage of correctly resolved conflicting pixels of only 31.52%. The learning parameters of the network were the same for all the tested patients. On average, 30,000 iterations and a learning rate of 1e-9 is sufficient in order to obtain accurate results. But these parameters could be adjusted according to the patient in order to avoid some patients having a bad conflict resolution and to improve the segmentation accuracy.

The more classical method of Corbat et al. [2] obtains the best results for conflict management, and the DL-OV^{2}ÅSSION method obtains the best segmentations. It would then be interesting to combine these two methods in order to obtain the best results in terms of conflict management and segmentation.

Currently, the presented patient segmentation process is completed in eight hours, and all the fusion methods presented are a process performed over several hours. The final execution times (segmentation and fusion) may seem long, but it must be taken into account that experts (surgeons and radiologists) are not monopolized by this calculation time. They can indeed devote themselves to other tasks and let the supercomputer carry out the segmentation and fusion. Moreover, we are not in a state of absolute emergency and experts believe that it is reasonable to produce a numerical representation of one patient's abdomen in a few days or even a week.

Finally, it should be noted that our new fusion method is, unlike the other methods, a semi-automatic method, in the sense that several slices of a patient must already be manually segmented to be used during training. In the SAIAD project, we have no problem with this, as we already have these slices. They are used in the special segmentation process, which is also semi-automatic. We cannot, at this time, perform automatic segmentation due to the unpredictability of the tumour and deformed kidney and the limited number of patients we can obtain.

Moreover, it should be noted that the training dataset used for our method is different to the training dataset of the other comparative networks. These different methods are therefore not strictly comparable. However, the training dataset of our method is a lot smaller, only 20% of a patient's slices are used for the training, compared to other fusion networks whose training dataset is composed of all the slices of 13 patients, and our method achieves a better fusion (compared to the other IA fusion methods) and improved the quality of the segmentations.

5. Conclusion and further work

In this paper, we have presented a synthesis of the work carried out in the SAIAD project. The segmentation method and fusion methods developed were tested on a larger set of patients for the representation of tumoral kidneys in children. In addition, a final neural network fusion system for complementary segmentation has been proposed, called DL-OVA^{2}SSION. This method allows for the resolution of conflicts and thus improves the accuracy of segmentation. This method of fusion increases the robustness of the general system of the SAIAD project.

Since we have observed that our previous method using six different mathematical methods was outperformed by other methods in the resolution of conflicting pixels, we will study the possibility of combining this fusion method, which obtained the best results for conflict management, with our DL-OV^{2}ASSION fusion method, which obtained the best results for segmentation. Further work will also focus on improving the system by adding functionalities for segmentation and fusion of new structures appearing on the scanners, such as arteries, veins, and urinary cavities.

Acknowledgements

The authors wish to thank Professor F. Auber and Dr M. Lenoir-Auber from the *BESANCON University Hospital* for their expertise in nephroblastoma, and Loredane Vieille and Elise Lacroix for performing the manual segmentations. The authors would like to thank the European Union for financing this project as part of the SAIAD INTERREGV program and the SAIAD consortium partners. Computations were performed on the supercomputer facilities of the Mésocentre de calcul de Franche-Comté. Finally, the authors also wish to thank Mary Moritz (WordStyle Traduction) for her help with the English language.

References

[1] Florent Marie, Lisa Corbat, Yann Chaussy, Thibault Delavelle, Julien Henriet, and Jean-Christophe Lapayre. Segmentation of deformed kidneys and nephroblastoma using case-based reasoning and convolutional neural network. *Expert Systems with Applications*, 127:282–294, 2019. `doi:10.1016/j.eswa.2019.03.010`.

[2] Lisa Corbat, Julien Henriet, and Jean-Christophe Lapayre. Conflict management in the fusion of complementary segmentations of deformed kidneys and nephroblastoma. *Medical Image Analysis*, 60, 2020. `doi:10.1016/j.media.2019.101629`.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436, 2015. `doi:10.1038/nature14539`.

[4] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. doi:10.1109/CVPR.2015.7298965.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. doi:10.1109/TPAMI.2016.2644615.

[6] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015. doi:10.1109/ICCV.2015.178.

[7] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. doi:10.1109/TPAMI.2017.2699184.

[8] William Thong, Samuel Kadoury, Nicolas Piché, and Christopher J Pal. Convolutional networks for kidney segmentation in contrast-enhanced ct scans. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, 6(3):277–282, 2018. doi:10.1080/21681163.2016.1148636.

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. doi:10.1007/978-3-319-24574-4_28.

[10] Xabier Artaechevarria, Arrate Munoz-Barrutia, and Carlos Ortiz-de Solórzano. Combination strategies in multi-atlas image segmentation:

application to brain mr data. *IEEE transactions on medical imaging*, 28(8):1266–1277, 2009. `doi:10.1109/TMI.2009.2014372`.

[11] Iván Cabria and Iker Gondra. Mri segmentation fusion for brain tumor detection. *Information Fusion*, 36:1–9, 2017. `doi:10.1016/j.inffus.2016.10.003`.

[12] Max Mignotte. A label field fusion model with a variation of information estimator for image segmentation. *Information Fusion*, 20:7–20, 2014. `doi:10.1016/j.inffus.2013.10.012`.

[13] Dac Cong Tai Nguyen, Said Benameur, Max Mignotte, and Frédéric Lavoie. Superpixel and multi-atlas based fusion entropic model for the segmentation of x-ray images. *Medical image analysis*, 48:58–74, 2018. `doi:10.1016/j.media.2018.05.006`.

[14] Max Mignotte and Charles Hélou. A precision-recall criterion based consensus model for fusing multiple segmentations. *Int J Signal Process Image Process Pattern Recognit*, 7(3):61–82, 2014. `doi:10.14257/ijsip.2014.7.3.07`.

[15] Lazhar Khelifi and Max Mignotte. A novel fusion approach based on the global consistency criterion to fusing multiple segmentations. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(9):2489–2502, 2016. `doi:10.1109/TSMC.2016.2531645`.

[16] Max Mignotte. A label field fusion bayesian model and its penalized maximum rand estimator for image segmentation. *IEEE Transactions on Image Processing*, 19(6):1610–1624, 2010. `doi:10.1109/TIP.2010.2044965`.

[17] Yuncong Feng, Xuanjing Shen, Haipeng Chen, and Xiaoli Zhang. Segmentation fusion based on neighboring information for mr brain images. *Multimedia Tools and Applications*, 76(22):23139–23161, 2017. `doi:10.1007/s11042-016-4098-3`.

[18] Yang Hu, Andrea Soltoggio, Russell Lock, and Steve Carter. A fully convolutional two-stream fusion network for interactive image segmentation. *Neural Networks*, 109:31–42, 2019. doi:10.1016/j.neunet.2018.10.009.

[19] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016. doi:10.1109/cvpr.2016.58.

[20] Youbao Tang, Xiangqian Wu, and Wei Bu. Deeply-supervised recurrent convolutional neural network for saliency detection. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 397–401. ACM, 2016. doi:10.1145/2964284.2967250.

[21] Youbao Tang and Xiangqian Wu. Saliency detection via combining region-level and pixel-level predictions with cnns. In *European Conference on Computer Vision*, pages 809–825. Springer, 2016. doi:10.1007/978-3-319-46484-8_49.

[22] Fen Xiao, Wenzheng Deng, Liangchan Peng, Chunhong Cao, Kai Hu, and Xieping Gao. Msdnn: Multi-scale deep neural network for salient object detection. *IET Image Processing*, 12(11):2036–2041, 2018. doi:10.1049/iet-ipr.2018.5631.

[23] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017. doi:10.1109/TIP.2017.2682981.

[24] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in neural information processing systems*, pages 109–117, 2011.

[25] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the*

22nd ACM international conference on Multimedia, pages 675–678, 2014. doi:10.1145/2647868.2654889.

[26] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015. doi:10.1007/s11263-014-0733-5.