



Learning anatomical digital twins in pediatric 3D imaging for renal cancer surgery

Giammarco La Barbera

► To cite this version:

Giammarco La Barbera. Learning anatomical digital twins in pediatric 3D imaging for renal cancer surgery. Artificial Intelligence [cs.AI]. Institut Polytechnique de Paris, 2022. English. NNT : 2022IPPAT040 . tel-03911159

HAL Id: tel-03911159

<https://theses.hal.science/tel-03911159>

Submitted on 22 Dec 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Learning anatomical digital twins in pediatric 3D imaging for renal cancer surgery

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à Télécom Paris

École doctorale n°626 Institut Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat: Signal, Images, Automatique et robotique

Thèse présentée et soutenue à Palaiseau, le 29 November 2022, par

Giammarco La Barbera

Composition du Jury :

Nicolas Passat Professeur, Université de Reims Champagne-Ardenne (CReSTIC)	Président
Olivier Bernard Professeur, Université de Lyon (CREATIS)	Rapporteur
Elena De Momi Professeure, Politecnico di Milano (NEARLAB)	Rapporteuse
Hervé Brisse Chef de service, Institut Curie (Radiology)	Examineur
Isabelle Bloch Professeure, Télécom Paris (LTCI) and Sorbonne Université (LIP6)	Directrice de thèse
Laurence Rouet Scientifique senior, Philips Research France	Co-directrice de thèse
Pietro Gori Maître de conférences, Télécom Paris (LTCI)	Co-directeur de thèse
Sabine Sarnacki Professeure et chef de service, Hôpital universitaire Necker-Enfants malades (Pediatric Visceral, Urological and Transplant Surgery)	Co-directrice de thèse

Learning anatomical digital twins in pediatric 3D imaging for renal cancer surgery

PhD manuscript of the Institut Polytechnique de Paris
prepared at Télécom Paris

Doctoral school n°626 Institut Polytechnique de Paris (ED IP Paris)
PhD Specialty: Signal, Images, Automatique et robotique

PhD manuscript presented and defended at Palaiseau, 29 November 2022, by

Giammarco La Barbera

Members of the jury:

Nicolas Passat Full professor, Université de Reims Champagne-Ardenne (CReSTIC)	President
Olivier Bernard Full professor, Université de Lyon (CREATIS)	Reviewer
Elena De Momi Full professor, Politecnico di Milano (NEARLAB)	Reviewer
Hervé Brisse Head of service, Institut Curie (Radiology)	Examiner
Isabelle Bloch Full professor, Télécom Paris (LTCI) and Sorbonne Université (LIP6)	Thesis supervisor
Laurence Rouet Senior scientist, Philips Research France (Ultrasound Image Formation and Applications)	Co-supervisor
Pietro Gori Assistant Professor, Télécom Paris (LTCI)	Co-supervisor
Sabine Sarnacki Full professor and head of service, Hôpital universitaire Necker-Enfants malades (Pediatric Visceral, Urological and Transplant Surgery)	Co-supervisor

Acknowledgements

I would like to use this opportunity to express my gratitude to everyone who was part of this journey that was my doctoral thesis.

First of all, I would like to thank the members of the jury who took the time to read, evaluate and provide feedback on this work, which led to a constructive and stimulating discussion. Your research work has been a great inspiration to me and I hope that mine can help you find new insights, and that our collaboration established with this jury can continue.

Then, I cannot thank my supervisors enough for all the help and support they have given me, each of you has taught me something that will always stay with me and you have helped me grow so much both humanly and scientifically. Thank you very much again, and I am very happy that I will continue to work with almost all of you and I hope to be able to collaborate with others as well, to further improve the work done and explore new horizons in the world of medical imaging processing.

I also want to thank all the other people who collaborated and made this thesis possible with their ideas and with their medical and scientific knowledge, with which we have managed to produce papers and articles of interest to the scientific community.

I thank all the people from the IMAGES team at Télécom Paris and the IMAG2 team at Necker hospital who made this journey even better, whose companionship both physically and remotely made it possible to get through tough times during the complicated pandemic.

I also thank the people in the FCC lab at Necker with whom we share the lab who make every day at the hospital even more enjoyable, and the group at Philips with whom I have unfortunately had little interaction but with whom every discussion has been of great help in this work.

I want to thank all the professors and engineers who also pushed, instructed and helped me on the worlds of computer science, biomedical engineering and medical image processing from high school to university.

Finally, I want to thank all my friends who have supported me over these 3 years, especially those in Paris who have always been there to cheer me up on difficult days and make good days even better. You make this city more of a home every day.

My family who despite the distance always managed to give me all they could, that made me the person that I am and who gave me the greatest gift by being present on the day of the defense. Grazie veramente tanto.

And then Alice, who has been my number one supporter in this adventure and to whom I dedicate this work. Words are not enough to thank you.

Merci beaucoup à vous tous encore une fois !

Abstract

Pediatric renal cancers account for 9% of pediatric cancers with a 9/10 survival rate at the expense of the loss of a kidney. Nephron-sparing surgery (NSS, partial removal of the kidney) is possible if the cancer meets specific criteria (regarding volume, location and extent of the lesion). Indication for NSS is relying on preoperative imaging, in particular X-ray Computerized Tomography (CT). While assessing all criteria in 2D images is not always easy nor even feasible, 3D patient-specific models offer a promising solution. Building 3D models of the renal tumor anatomy based on segmentation is widely developed in adults but not in children. There is a need of dedicated image processing methods for pediatric patients due to the specificities of the images with respect to adults and to heterogeneity in pose and size of the structures (subjects going from few days of age to 16 years). Moreover, in CT images, injection of contrast agent (contrast-enhanced CT, ceCT) is often used to facilitate the identification of the interface between different tissues and structures but this might lead to heterogeneity in contrast and brightness of some anatomical structures, even among patients of the same medical database (i.e., same acquisition procedure). This can complicate the following analyses, such as segmentation.

The first objective of this thesis is to perform organ/tumor segmentation from abdominal-visceral ceCT images. An individual 3D patient model is then derived. Transfer learning approaches (from adult data to children images) are proposed to improve state-of-the-art performances. The first question we want to answer is if such methods are feasible, despite the obvious structural difference between the datasets, thanks to geometric domain adaptation. A second question is if the standard techniques of data augmentation can be replaced by data homogenization techniques using Spatial Transformer Networks (STN), improving training time, memory requirement and performances.

In order to deal with variability in contrast medium diffusion, a second objective is to perform a cross-domain CT image translation from ceCT to contrast-free CT (CT) and vice-versa, using Cycle Generative Adversarial Network (CycleGAN). In fact, the combined use of ceCT and CT images can improve the segmentation performances on certain anatomical structures in ceCT, but at the cost of a double radiation exposure. To limit the radiation dose, generative models could be used to synthesize one modality, instead of acquiring it. We present an extension of CycleGAN to generate such images, from unpaired databases. Anatomical constraints are introduced by automatically selecting the region of interest and by using the score of a Self-Supervised Body Regressor, improving the selection of anatomically-paired images between the two domains (CT and ceCT) and enforcing anatomical consistency.

A third objective of this work is to complete the 3D model of patient affected by renal tumor including also arteries, veins and collecting system (i.e. ureters). An extensive study and benchmarking of the literature on anatomic tubular structure segmentation is presented. Modifications to state-of-the-art methods for our specific application are also proposed. More-

over, we present for the first time the use of the so-called vesselness function as loss function for training a segmentation network. We demonstrate that combining eigenvalue information of the Hessian matrix of segmentation masks with structural and voxel-wise information of other loss functions results in an improvement in performance.

Eventually, a tool developed for using the proposed methods in a real clinical setting is shown as well as a clinical study to further evaluate the benefits of using 3D models in pre-operative planning. The intent of this research is to demonstrate through a retrospective evaluation of experts how criteria for NSS are more likely to be found in 3D compared to 2D images. This study is still ongoing.

Résumé

Les cancers rénaux pédiatriques représentent 9% des cancers pédiatriques avec un taux de survie de 9/10 au prix de la perte d'un rein. La chirurgie d'épargne néphronique (NSS, ablation partielle du rein) est possible si le cancer répond à des critères précis (e.g. le volume et la localisation de la lésion). L'indication de la NSS repose sur l'imagerie préopératoire, en particulier la tomographie informatisée à rayons X (CT). Si l'évaluation de tous les critères sur des images 2D n'est pas toujours facile, les modèles 3D spécifiques au patient offrent une solution prometteuse. La construction de modèles 3D de l'anatomie rénale basés sur la segmentation est développée chez les adultes mais pas chez les enfants. Il existe un besoin de méthodes de traitement d'image dédiées aux patients pédiatriques en raison des spécificités de ces images, comme l'hétérogénéité de la pose et de la taille des structures. De plus, dans les images CT, l'injection d'un agent de contraste est souvent utilisée (ceCT) pour faciliter l'identification de l'interface entre les différentes structures mais cela peut conduire à une hétérogénéité dans le contraste de certaines structures anatomiques, même parmi les patients acquis avec la même procédure.

Le premier objectif de cette thèse est d'effectuer une segmentation des organes/tumeurs à partir d'images ceCT, à partir de laquelle un modèle 3D sera dérivé. Des approches d'apprentissage par transfert (des données adultes aux images enfants) sont proposées. La première question consiste à savoir si de telles méthodes sont réalisables, malgré la différence structurelle évidente entre les ensembles de données. Une deuxième question porte sur la possibilité de remplacer les techniques standard d'augmentation des données par des techniques d'homogénéisation des données utilisant des Spatial Transformer Networks, améliorant ainsi le temps d'apprentissage, la mémoire requise et les performances.

La segmentation de certaines structures anatomiques dans des images ceCT peut être difficile à cause de la variabilité de la diffusion du produit de contraste. L'utilisation combinée d'images CT sans contraste (CT) et ceCT atténue cette difficulté, mais au prix d'une exposition doublée aux rayonnements. Le remplacement d'une des acquisitions CT par des modèles génératifs permet de maintenir la performance de segmentation, en limitant les doses de rayons X. Un deuxième objectif de cette thèse est de synthétiser des images ceCT à partir de CT et vice-versa, à partir de bases d'apprentissage d'images non appariées, en utilisant une extension des Cycle Generative Adversarial Networks. Des contraintes anatomiques sont introduites en utilisant le score d'un Self-Supervised Body Regressor, améliorant la sélection d'images anatomiquement appariées entre les deux domaines et renforçant la cohérence anatomique.

Un troisième objectif de ce travail est de compléter le modèle 3D d'un patient atteint d'une tumeur rénale en incluant également les artères, les veines et les uretères. Une étude approfondie et une analyse comparative de la littérature sur la segmentation des structures tubulaires anatomique sont présentées. En outre, nous présentons pour la première fois l'utilisation de la fonction "vesselness" comme fonction de perte pour l'entraînement d'un réseau de segmen-

tation. Nous démontrons que la combinaison de l'information sur les valeurs propres de la matrice hessienne des masques de segmentation avec les informations structurelles d'autres fonctions de perte permet d'améliorer les performances.

Enfin, nous présentons un outil développé pour utiliser les méthodes proposées dans un cadre clinique réel ainsi qu'une étude clinique visant à évaluer les avantages de l'utilisation de modèles 3D dans la planification préopératoire. L'objectif de cette recherche est de démontrer, par une évaluation rétrospective menée par des experts, comment les critères du NSS sont plus susceptibles d'être trouvés dans les images 3D que dans les images 2D. Cette étude est toujours en cours.

Summary

1	Introduction	15
1.1	Medical context	15
1.2	Goals, contributions and questions	19
1.3	Organization of the manuscript	20
2	The Necker PRAC database	21
2.1	Database creation	21
2.2	Necker CT acquisition protocol	22
2.3	Pediatric dataset selection and specifications	23
3	Segmentation of kidneys and renal tumors on pediatric abdominal-visceral ceCT scanners via deep learning	27
3.1	Related work	27
3.1.1	Overview	27
3.1.2	No-new-U-Net	28
3.2	Application of no-new-U-Net on a database of adult images	32
3.2.1	Adults database of KiTS19 Challenge	32
3.2.2	Results with the KiTS19 dataset from using and rebuilding nnU-Net	33
3.3	Transfer learning from adults to children	34
3.4	Homogenization with Spatial Transformers	37
3.4.1	Proposed method	38
3.4.2	Results and discussion	41
3.5	Transfer learning with <i>common</i> size and pose	47
3.6	Conclusion	49
4	Cross-domain CT image translation using CycleGAN	51
4.1	Anatomical constrained CycleGAN	54
4.1.1	Input selection via SSBR	55
4.1.2	Anatomically constrained CycleGAN	57
4.2	Results and discussion	59
4.2.1	Implementation details	59
4.2.2	Qualitative results on unpaired datasets	60
4.2.3	Quantitative ablation study on paired database	68
4.2.4	Blood vessel segmentation using ceCT and CT	68
4.3	Conclusion	73

5	Segmentation of tubular structures on pediatric abdominal-visceral ceCT scanners with renal tumor	75
5.1	Assessment of State-Of-The-Art methods	77
5.1.1	Vesselness filters	77
5.1.2	Rule-based methods for tubular structures segmentation	79
5.1.3	Deep learning-based methods for tubular structures segmentation	80
5.1.4	Assessment considerations	83
5.2	Methods	83
5.2.1	Comparison of state-of-the-art methods of renal tubular structures segmentation in ceCT images	83
5.2.2	Proposed tubular structures loss function	85
5.3	Materials and Experiments	90
5.3.1	Database	90
5.3.2	Training implementation details	91
5.3.3	Evaluation measures	92
5.4	Results and Discussion	92
5.4.1	Arteries and veins segmentation	92
5.4.2	Ureters segmentation	97
5.5	Conclusions	99
6	Application of anatomical digital twins for renal cancer surgery	101
6.1	Preparation of anatomical models via a 3DSlicer plug-in	101
6.1.1	From ceCT scan to 3D volume: the “Renal Anatomy Segmentation For ceCT” module for 3DSlicer	101
6.1.2	From 3DSlicer to the operating room: an anatomical digital twin to help surgeons	105
6.2	Advantages of 3D models in pre- and per-operative planning	105
6.2.1	Analysis of two interesting clinical cases	106
6.2.2	Retrospective on-going study on 3D model vs. 2D imaging	109
6.3	Conclusion	111
7	Conclusions and perspectives	113
7.1	Conclusions	113
7.2	Perspectives	119
	Publications	125
	A Evaluation measures	127
	B Parameters details	129
	C Supplementary material for kidney and renal tumor segmentation	131
	D Quantitative results on state-of-the-art methods for ceCT-CT translation	135
	E Supplementary material for segmentation of tubular structures	137
	F Details of the “Renal Anatomy Segmentation for ceCT” module	145

SUMMARY

13

Bibliography

153

Chapter 1

Introduction

1.1 Medical context

Pediatric renal cancers account for 8-9% of pediatric cancers and are mainly nephroblastoma (95%, also called Wilms' tumors - WT) [58]. They are common in children younger than 5 years old but 9/10 survive at the expense of the loss of a kidney. As a matter of fact, when one kidney is removed, the other one takes over the full job of filtering, resulting in a higher risk of long-term cardiovascular disease such as hypertension, and of chronic renal insufficiency. According to the new international treatment protocol (Umbrella SIOP Protocol [12]), radical nephrectomy (RN, total kidney removal) is the current standard for unilateral tumors, while nephron-sparing surgery (NSS, partial removal) is recommended in some specific cases such as bilateral WT (5% of the cases), syndromic patients with a predisposition to bilateral tumors occurring, and unilateral WT with a diseased contralateral kidney. In the last years it has also been proposed in unilateral non-syndromic WT [105] if specific criteria are met (concerning volume, location and extent of the lesion within and out of the kidney, as well as estimated amount of the remaining spared tissue). The evaluation of this conservative approach is one of the secondary aims of the Umbrella SIOP protocol. Indication for NSS relies on preoperative imaging (after four weeks of neoadjuvant chemotherapy), and X-ray contrast-enhanced Computerized Tomography (ceCT) is usually done. Another important indication for medical doctors on the SIOP protocol is how to decide the surgery approach, whether laparoscopic (LS) or open (laparotomic, LT) approach. Even if laparoscopic NSS is not yet recommended. Assessing all protocol criteria in 2D images is not always easy nor even feasible, and 3D models could be a good solution. Moreover, these models could help surgeons as per-operative guidance. An abdomino-visceral anatomical 3D model should include kidneys, vertebrae, ribs, arteries, veins and tumors; ureters are included if the injection makes them visible, while spleen and liver if the tumor makes contact with them. An example is shown in Figure 1.1.

Building a 3D model of renal tumor anatomy based on image segmentation of ceCT images is now becoming popular [36, 59, 109]. Automatic segmentation methods speed up model creation procedures and reduce the inter-subject variability in comparison to manual segmentation. While such methods are highly developed [53, 84, 102] and perform well for some structures such as kidneys and tumors for adults [63, 95], this is not true for children. There is a need of dedicated image processing methods for pediatric patients due to the specificities of the images with respect to adults (different contrast, size, etc.), to the anatomical and topological changes related to development, and to specific pediatric pathologies. To address



Figure 1.1: Example of a partial 3D model of a pediatric patient affected by nephroblastoma who presents different tumors. The structures represented are kidneys (in brown), arteries (red), veins (blue), ureters (gray), spleen (dark blue), vertebrae and ribs (light gray), tumors (other colors). Left: Anterior view. Right: Posterior view without ribs.

this need, this thesis focuses on generating 3D anatomical models (digital twins) of pediatric renal cancer patients as automatically as possible through the use of deep learning algorithms. The use of such machine learning techniques makes it possible both to speed up the creation of the models due to their computational efficiency (leveraging GPUs) and to minimize the manual-interaction required from the physician. These advantages allow an increase in the number of digital twins that a hospital laboratory can prepare in a workday and save physician valuable time.

Nevertheless, developing machine learning algorithms, and especially deep learning ones, for segmenting pediatric images is a challenging task. First, pediatric datasets contain subjects ranging from few days of age to 16 years, showing therefore anatomical structures which are *highly heterogeneous in terms of size*. Furthermore, due to the fact that children do not always stand still during the acquisition [35], pediatric images also present a *high variability in terms of pose* and movements artifacts. Figure 1.2 (left) illustrates these problems.

Moreover, pediatric databases are limited in number of images [67] and therefore usual deep learning strategies might fail or might not give good results [126]. Direct inference or transfer learning from networks trained on adults might fail because of the differences between the two populations, especially in terms of relative size between organs and variability among subjects [142] (Figure 1.2 right). Some authors proposed to use an ad-hoc data augmentation method, as in [77], to take into account the differences between adults and children. However, this usually takes time, and it is not always possible nor easy to recreate all the sources of variations (e.g. relative size between organs and tumors) in such a data augmentation process. Another important aspect is the heterogeneity in image intensity among ceCT images and in a same individual ceCT image, in particular in the abdominal-visceral region. Pixel intensities after contrast agent injection do not always change equally from patient to patient, and even in different regions of one patient, due to different factors, such as the presence of the tumor, thrombosis in the vessels or simply a different acquisition time. These factors lead to *inter-patient* heterogeneity in the contrast and brightness of the same structures among

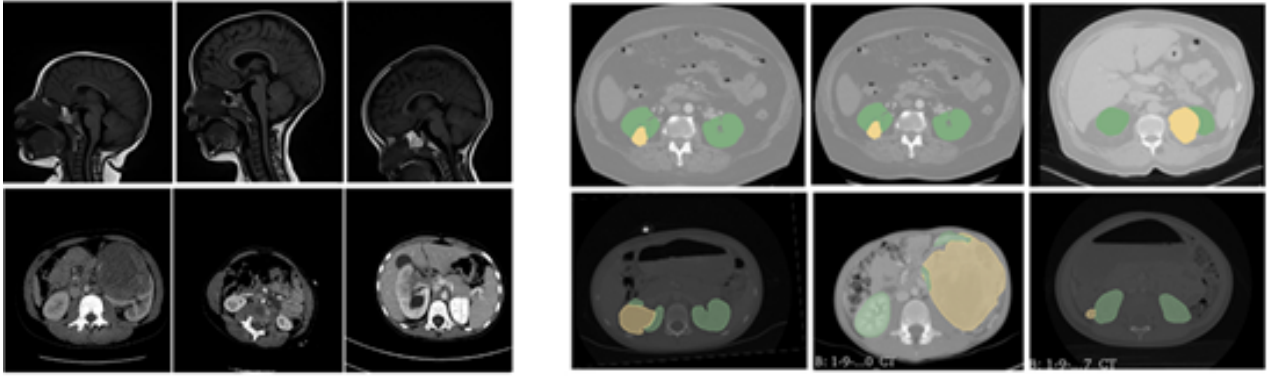


Figure 1.2: Left: Differences in size and pose among patients in two of our pediatric datasets. First row: MR sagittal brain images gathered at Bicetre Hospital of Paris for a previous work of mine [77]. Second row: CT axial abdominal images from the Necker PRAC database (see Chapter 2). Right: Differences between MICCAI KiTS19 [53] adult images (first row) and Necker PRAC pediatric abdominal images (second row). Kidneys are in green and tumors in yellow.

patients of the same medical database. Furthermore, due to the above mentioned reasons, contrast medium may not reach some parts of the same structure, causing *intra-patient* contrast variability within an anatomical structure. Some examples of differences in contrast medium diffusion are shown in Figure 1.3.

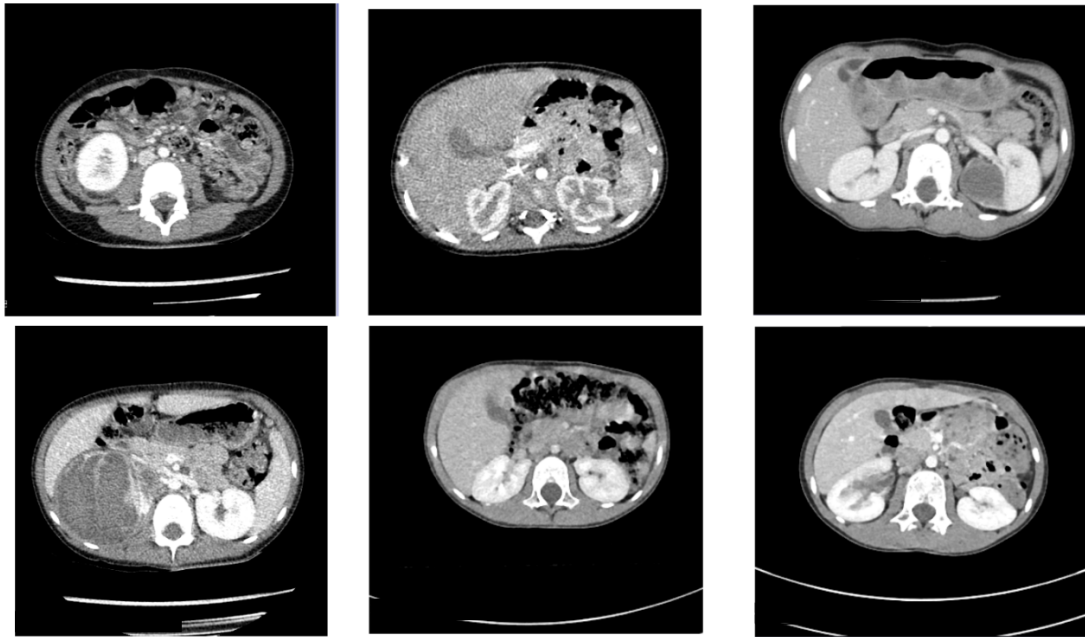


Figure 1.3: Some examples of arteriovenous phase ceCT images from the paediatric and pathological dataset of Necker Hospital.

Some early researches [119, 125, 158] have demonstrated that the combined use of ceCT and contrast-free (CT) CT images can tackle the variability in contrast medium diffusion and improve the segmentation performances, but at the cost of a double radiation exposure which radiologists, in particular in pediatrics, prefer to avoid. To limit the radiation dose, generative models could be used to synthesize one modality, instead of acquiring it. The Cycle Generative Adversarial Network (CycleGAN) [157] approach has recently attracted particular attention

because it alleviates the need for paired data that are difficult to obtain, even if it is still far from achieving clinical acceptance level. In fact, despite the great performances demonstrated in the literature for some tasks, limitations [25, 121, 149] still remain when dealing with 3D volumes generated slice by slice from unpaired datasets with different fields of view.

Renal tubular structures, such as ureters, arteries and veins, are most affected by intra- and inter-patient contrast heterogeneity, and it can be challenging to segment them via deep learning algorithms. Nevertheless, they are very important for building a 3D digital twin of the patient that is as complete as possible. In fact, the surgeon has to control the flow of blood into and out of the kidney, which is fundamental to avoid ischemia or hemorrhaging but also to preserve vascularization when a NSS is performed. Additionally the preservation of the collecting system, namely ureters, allows correct urine flow and preserve long-term renal function. A digital twin with all the renal tubular structures segmented helps in both tasks, allowing for vasculature preservation and easier recognition of accessory renal blood vessels (multiple vessels are present in 35% of the patients), and for a better understanding of the renal pelvis (including the attachment of the ureter to the kidney) [1]. For abdominal-visceral ceCT images, radiologists opt for early arterial phase acquisition to achieve high contrasted arteries or late delayed phase acquisition for high contrasted excretory pathways (i.e. ureters). Examples of each phase acquisition time in relation to contrast agent injection in CT are displayed in Figure 1.4.

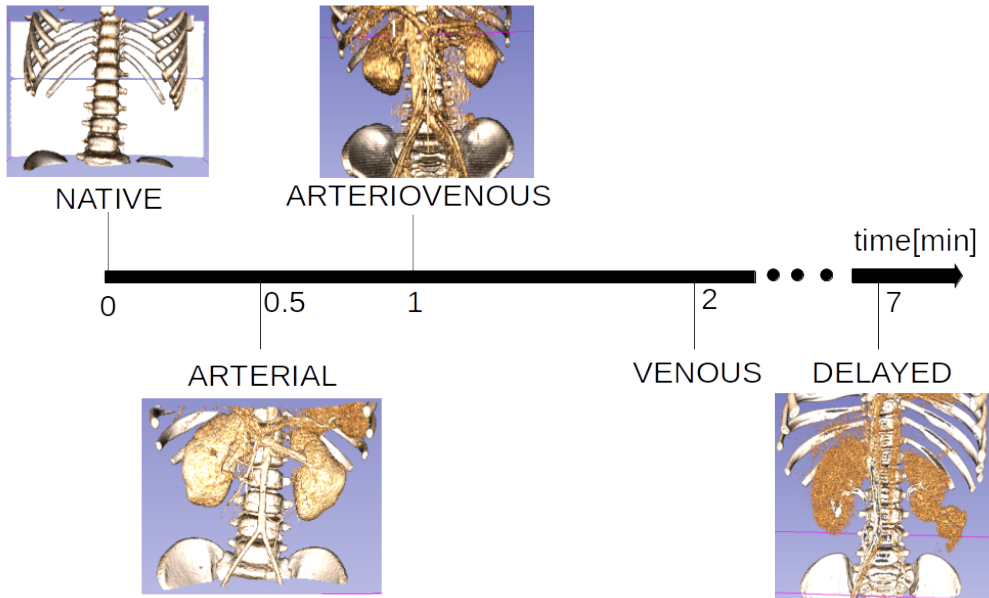


Figure 1.4: Exemplary timestamps of each phase acquisition in relation to the contrast agent injection. Adapted from [54] using volume-rendered images from Necker PRAC dataset.

Nevertheless, due to the very rapid occurrence of the arterial phase (in particular in children who have very rapid blood circulation given a higher heart rate) or due to medical choices in order to have both arteries and veins with higher intensity (preferred in renal tumor cases), very often ceCT images are acquired in a late arterial phase, also known as arteriovenous phase [54]. Here, both types of vessels are contrasted but the presence of less contrast medium in each one results in lower intensity (and thus less difference with other structures) as well as greater heterogeneity (which may be accentuated by the presence of tumor or thrombus).

This increases the difficulties mentioned above, which are joined by the challenge of segmenting structures with elongated shape and intra-scale changes. Some authors (e.g. [51, 85, 128]) tried to solve these issues on adults abdominal ceCT images without great results or limited to specific acquisition CT modalities. The difficulties of segmenting such tubular structures are also increased in pediatric subjects, due to inter-anatomy variation, small volume to background ratio (where by background we mean everything that it is not the structures of interest), and small available labeled dataset.

1.2 Goals, contributions and questions to be answered

The final goal of this work is to create pre- and per-operative individual 3D anatomical models from pediatric abdominal-visceral ceCT scanners with renal tumors to help surgeons verify the criteria to proceed with a NSS and to guide the surgery. In order to achieve this aim, while at the same time minimizing the manual interaction required from the clinician, the first goal of this thesis is the automatic segmentation of such images using deep learning techniques. To this end, four contributions are made: three to achieve the first goal and a fourth as an intermediate step to reach the second and final goal.

The first contribution of this thesis is the kidneys and renal tumor segmentation from the pediatric ceCT images gathered at Necker hospital of Paris (see Necker PRAC database in Chapter 2). Transfer learning approaches (from adult data to children images, based on existing pre-trained weights on adults) are proposed to improve state-of-the-art performances. The questions we want to answer are whether such methods are possible despite the obvious structural differences between the datasets, and whether the standard techniques of data augmentation can be replaced by data homogenization techniques, improving training time, memory required and performance.

The second contribution of this work is the segmentation of renal tubular structures (namely as arteries, veins and ureters) which is divided into two propositions.

First, in order to deal with the variability in contrast medium diffusion of contrast-enhanced CT, that most affects renal tubular structures, we propose the segmentation using both ceCT and contrast-free CT during training, combined with iconographic data augmentation. However, due to the presence of only one modality in real-setting CT databases, as in the Necker PRAC database and namely ceCT, generative models are used to synthesize the missing modality (i.e. CT). We want to understand if it is possible to leverage anatomical constraints, to generate high fidelity images in ceCT-CT translation (both directions), with better structural consistency than state-of-the-art methods. In addition, we want to examine whether the use of such anatomical constrained synthetic images allows us to achieve performance on par with the segmentation performance using both real modalities.

Secondly, in order to find inspirational approaches, the state-of-the-art methods for renal tubular structures on adults are assessed due to the lack of literature in children. The questions we want to answer are whether such methods perform well also on the pediatric and pathological ceCT images with arteriovenous phase, and whether the standard voxel-wise loss functions can be combined with a new loss function designed for tubular structures and based on their morphological information, improving both spatial and structural performances.

Once the first goal is achieved (entirely or partially due to possible difficulties encountered) and in order to achieve the final one, our fourth contribution is the application on a real clinical setting of the proposed methods based on deep learning to perform automatic segmentation. In

order to make such approaches usable for physicians, a user-friendly software is implemented. We want to understand whether using this tool really speeds up the annotation needed to create the 3D anatomical model and reduces manual annotations. In addition, we want to examine the advantages of using digital twin versus using structural images alone for pre-operative planning and per-operative guidance on clinically relevant cases.

The contributions and goals are summarized in Figure 1.5.

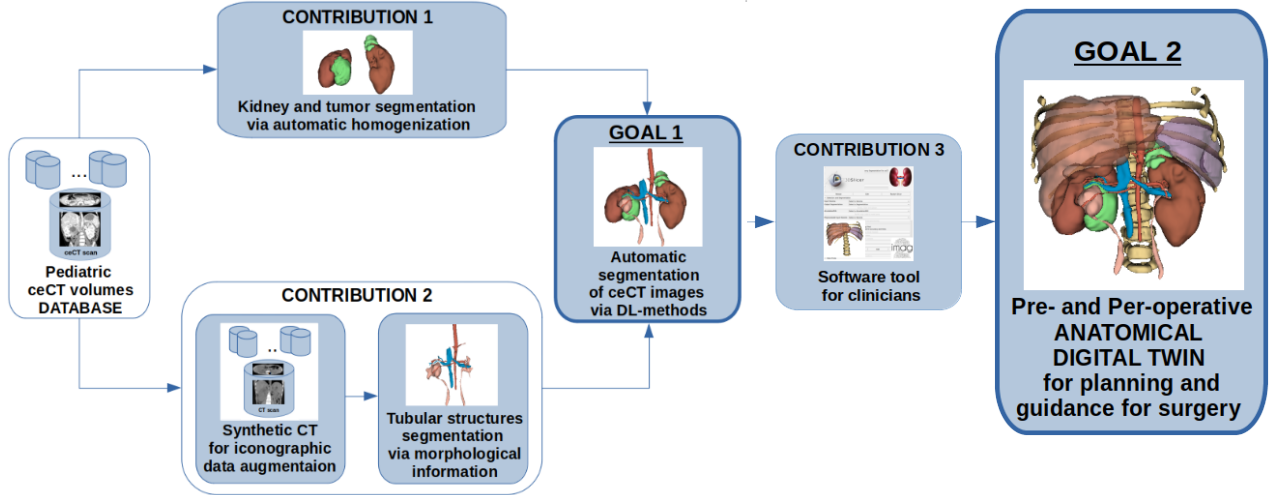


Figure 1.5: Contributions and goals of this PhD thesis. The ceCT database is presented in Chapter 2, while the contributions are presented from Chapter 3 to 6.

1.3 Organization of the manuscript

This PhD manuscript is structured as follows. Chapter 2 introduces and details the pediatric and pathological (namely renal tumor) abdominal-visceral ceCT database gathered at Necker hospital of Paris, whose 3D modeling via automatic segmentation is the core of this thesis.

In Chapter 3, we focus on kidney and renal tumor segmentation on ceCT scanners which show high differences in tumor development between adults and children (Figure 1.2 right). In addition, the use of extensive data augmentation is compared with a new technique for automatic image homogenization. Chapter 4 describes the application of a cycle generative adversarial model to obtain CT images from ceCT (and vice versa) for the purpose of data augmentation to tackle inter- and intra- patient contrast variability. In order to obtain anatomically consistent synthetic images an extension of the state-of-the-art technique for ceCT-CT translation is proposed. In Chapter 5, we present our assessment and comparison of state-of-the-art methods for tubular structures segmentation on the Necker PRAC database. In addition, a new loss function specifically adapted to tubular structures is proposed.

Chapter 6 describes the software tool designed in order to apply the method proposed for the creation of an anatomical 3D model in a real clinical setting. Furthermore, advantages of the use of the 3D digital twins are discussed, showing some interesting clinical cases to further evaluate these benefits. I personally consider this chapter to be of great importance, as it provides insight into the real value of the proposed approaches through their application for a day-to-day clinical use.

Chapter 7 summarizes and discusses the contributions of this thesis work.

Chapter 2

The Pediatric RenAl Cancer (PRAC) database of Necker hospital

In this chapter we present the abdominal-visceral ceCT pediatric and pathological (i.e. renal cancer) database gathered at the Necker hospital of Paris¹. Throughout this manuscript we refer to this database as the Necker PRAC (Pediatric RenAl Cancer) Database.

2.1 Database creation

The database was collected from PACS archive of Necker Hospital, for 115 patients (66 females and 49 males) from 3 days old to 18 years old. 49 patients have a left Wilms Tumor (WT), 48 a right WT, 11 a bilateral WT and 7 suffer of syndromes which could develop WT. Some patients have undergone more than one CT scanner exam (before or after chemotherapy, before or after surgery, or with and without contrast agent injection), so the number of available images is greater than the number of patients. These exams were performed in the course of the normal care pathway of the patient and were studied retrospectively after anonymization.

The creation of this database was a fundamental work for this thesis. The database is organized according to the following format:

```
X-J-WW (patient's code in Necker Database)
├── YYYY-MM-DD (exams per date)
│   ├── 3D
│   ├── MRI
│   ├── photos
│   ├── US
│   └── CT
│       ├── NotInjected
│       └── Injected
│           ├── DICOM
│           └── Processing
│               ├── Image in NIfTI format
│               └── Manual segmentation in NIfTI format
```

¹Hôpital Necker Enfants-Malades: service of Pediatric Visceral, Urological and Transplant Surgery, head of service: Pr Sabine Sarnacki; service of Pediatric Radiology, head of service: Pr Nathalie Boddaert.

where for patient's code X is the hospital, J the disease and WW the patient number (for example 1-9-10, X = 1 = Necker, J = 9 = Wilms' Tumor, WW = 10 = patient number 10 with WT), and for exams per date Y stands for the year, M for the month and D for the day (e.g. 2011-11-02 is November 2nd of 2011).

A reference Excel table was also created containing this information:

- **in the first sheet:** DeePRAC_ID (X-Y-WW), age (in years), exam date (YYYY-MM-DD), sex, localization of the tumor, CT operation time (Pre-Op or Post-Op), CT type, Injection (Yes or No), Notes (for additional information), Magnetic Resonance (MR) images (if also this exam is available), UltraSound (US) images (availability) and then the different structures to be segmented for every subject, which will be marked with an "X" once done;
- **in the second sheet:** Correspondence between DeePRAC_ID and Name of the patient. This sheet is strictly confidential and it is only found in the laboratory computer at Necker Hospital.

2.2 Necker CT acquisition protocol

There are three types of acquisition of contrast-enhanced CT:

- **vascular/early phase of mono-phasic injection:** in this phase we summarize both the purely arteriovenous phase and the arterial phase which often results also in an early arteriovenous phase, due to the problem mentioned previously in Introduction. In the arterial phase the image is acquired 30 seconds after injection (see Figure 1.4 in Introduction) to have the enhancement of the renal parenchyma and vasculature (even if only external parenchyma is well-contrasted, arteries are better contrasted than veins, and little renal masses could not be very visible but rupture in big tumor capsules are better visible) [153]. The arteriovenous phase, also called nephrographic phase, is acquired 50 seconds after injection (see Figure 1.4 in Introduction) to have the enhancement of renal parenchyma, blood vessels and others organs in proximity such as liver and spleen (external and internal parenchyma, as well as arteries and veins, are enhancement in a similar way, and little renal masses are easier detected) [153]. This is the phase usually acquired in the case of renal tumor, if only one acquisition is made.
- **excretory/delayed phase of mono-phasic injection:** the image is acquired 6 minutes after injection to have the enhancement of the excretory pathways (calyces, renal pelvises, and ureters, but the vasculature as well as the renal parenchyma are not well enhanced [153]), as shown in Figure 1.4 in Introduction.
- **bi-phasic injection:** this technique is instead composed by a unique acquisition acquired after two injections: a first one (with 1/3 of total amount of contrast medium) to have a excretory/delayed phase and the other (with 2/3 of contrast) to have a nephrographic phase. The second injection is performed after 6 minutes from the first one, then a unique image is acquired after another 30 or 50 seconds. This allows for the enhancement of both the renal parenchyma and all tubular structures. However we will have a less strong contrast enhancement [152] (see Figure 2.1), arteries and veins may

be difficult to distinguish between each other. The bi-phasic injection protocol started in April 2020 so only very few cases are available at the moment.



Figure 2.1: Volume-rendered example of single ceCT acquisition with bi-phasic injection in order to have both arteriovenous and delayed phases.

Examples of ceCT images with vascular phase, excretory phase and bi-phasic injection with respectively blood vessels, ureters and all structured labeled are shown in Figure 2.2.

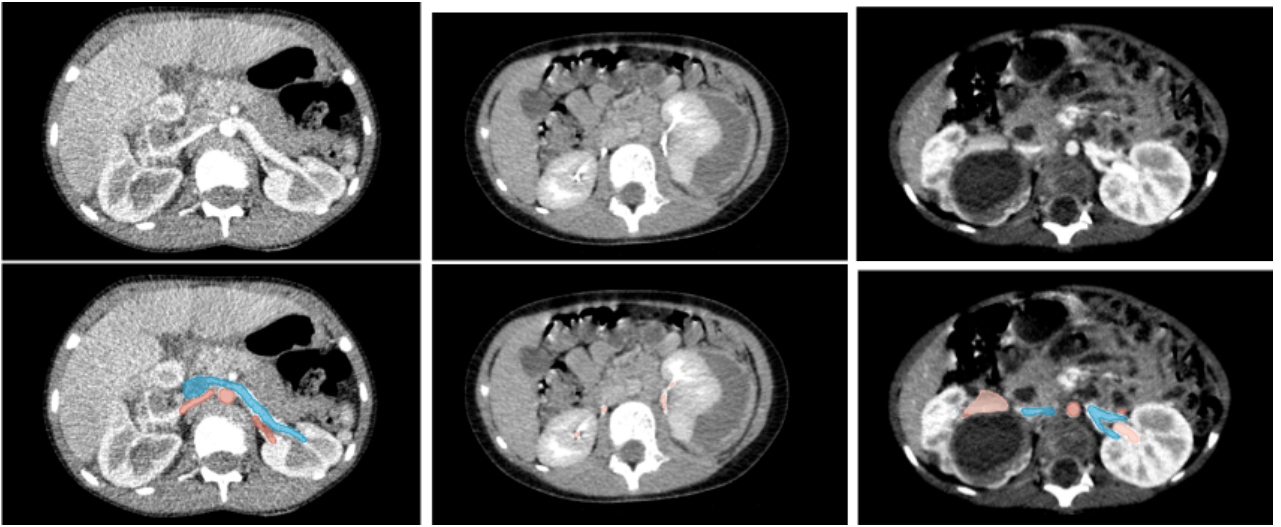


Figure 2.2: From left to right: ceCT images with vascular/early phase of mono-phasic injection, with excretory/delayed phase of mono-phasic injection and with bi-phasic injection. First row: original images. Second row: labeled images, with arteries in red, veins in blue and ureters in pink.

Some images in the database have been acquired with an old protocol (until 8 years ago) in which also an initial contrast-free (CT) CT was acquired to detect calcium or fat in a lesion and to provide baseline attenuation of any renal mass.

2.3 Pediatric dataset selection and specifications

Dataset selection. We decided to mainly work with pre-operative ceCT with mono-phasic injection of contrast medium, both at early phase and delayed phase, because these are highly

more numerous than bi-phasic ones. We selected a total of **80 volumetric images**, including 10 of these (vascular/early phase) that present also a contrast-free CT images.

Reference segmentations were performed by manual annotation under the supervision of medical experts using the open-source software 3DSlicer (<http://www.slicer.org>) [37]. After a training on abdominal-visceral anatomy by the surgeons and some side-by-side segmentation, manual annotations were made by me from scratch, and subsequently corrected first by the radiologists and then, in some cases, by the surgeons who participated in this work.

Specifically, the data were used for their characteristics in the following way:

- for **kidneys and tumors segmentation** (mainly used in Chapter 3) we used all the 80 ceCT scanners with both vascular/early and excretory/delayed phase of mono-phasic injection;
- for **arteries and veins segmentation** (used in both Chapter 4 and Chapter 5) we selected 63 ceCT scanners with vascular/early phase of mono-phasic injection and 3 with bi-phasic injection;
- for **ureters segmentation** (mainly used in Chapter 5) we selected 13 ceCT scanners with excretory/delayed phase of mono-phasic injection and 3 with bi-phasic injection;
- for the **quantitative evaluation of CycleGAN methods** (used in Chapter 4) we used the 10 patients with *paired* ceCT and CT scanners.

Other public databases used throughout this thesis work will be presented in the individual chapters in which they are used.

Database specifications. Some other details about volumes and contrast intensity of each structure are collected in Table 2.1 for training sets and Table 2.2 for test sets. The volumes are expressed in *ml*, while the image intensity represents the density in CT and is calculated using a linear attenuation coefficient expressed in Hounsfield units (HU) as follows:

$$\mu(HU) = 1000 \frac{\mu - \mu_{H_2O}}{\mu_{H_2O}} \quad (2.1)$$

where μ represents the linear attenuation coefficient (the fraction of the X-rays beam that is absorbed or scattered per unit thickness of the absorber tissue), and μ_{H_2O} represents the linear attenuation coefficient of the water. This densitometric scale is adimensional and traditionally includes 4000 values, starting from a minimum of -1000 , the region of ray transparent (e.g. air). In the higher extreme we find highly radiopaque elements such as bones, while obviously 0 HU corresponds to water density [20].

In the first and second columns of both Tables 2.1 and 2.2 we can notice a high variability in tumor volume, while in the third and forth columns we can observe a significant variability in HU for almost all the structures, with very high standard deviations for arteries, veins and collecting system. Heterogeneity in HU is slightly attenuated with the use of normalization (last two columns), namely a clipping of the intensity values to the 0.5 and 99.5 percentiles of the voxels of the structure under examination and a z-scoring normalization.

Images have been acquired with Siemens or GE Medical Systems CT Scan, and they have the following characteristics:

Table 2.1: Training sets. Normalization (HU_N): clipping in range [0.5,99.5] percentile and z-scoring.

Structure	Volume mean (std) ml	Volume min:max ml	Voxel value mean (std) HU	Voxel value min:max HU	Voxel value mean (std) HU_N	Voxel value min:max HU_N
Kidney Sx	67.52 (36.18)	12.06:235.94	166.46 (40.40)	69.33:293.16	1.31 (0.57)	-0.11:2.82
Kidney Dx	69.49 (42.18)	13.76:238.22	168.53 (42.19)	83.60:275.20	1.33 (0.59)	0.10:2.66
Tumor Sx	480.23 (583.16)	0.87:2704.02	60.21 (21.69)	18.61:117.95	-0.26 (0.32)	-0.85:0.61
Tumor Dx	314.67 (406.89)	6.13:2231.39	58.79 (20.09)	24.91:98.64	-0.27 (0.29)	-0.76:0.32
Arteries	9.72 (7.81)	1.54:46.13	205.37 (50.47)	121.87:431.24	0.09 (0.29)	-0.37:1.38
Veins	11.63 (14.75)	0.99:91.57	194.97 (216.30)	99.33:1614.95	-0.07 (0.57)	-0.51:3.32
Collecting System	3.79 (1.88)	1.10:7.76	1220.86 (557.37)	513.37:2485.9	-0.04 (0.62)	-0.82:1.36

Table 2.2: Test sets. Normalization (HU_N): clipping in range [0.5,99.5] percentile and z-scoring.

Structure	Volume mean (std) ml	Volume min:max ml	Voxel value mean (std) HU	Voxel value min:max HU	Voxel value mean (std) HU_N	Voxel value min:max HU_N
Kidney Sx	69.38 (24.15)	33.27:107.13	174.43 (47.76)	101.81:304.11	1.39 (0.58)	0.37:2.62
Kidney Dx	56.43 (26.18)	9.59:104.79	177.65 (60.83)	88.31:368.72	1.39 (0.64)	0.18:3.03
Tumor Sx	338.30 (405.52)	3.86:977.38	57.76 (15.22)	34.49:87.08	-0.28 (0.22)	-0.63:0.15
Tumor Dx	352.47 (582.81)	4.97:1944.85	48.41 (14.27)	32.22:81.67	-0.41 (0.21)	-0.66:0.08
Arteries	9.22 (4.75)	2.37:16.54	201.77 (40.63)	124.97:271.36	0.07 (0.23)	-0.36:0.47
Veins	13.91 (9.41)	3.52:29.99	163.61 (32.95)	118.54:233.6	-0.14 (0.19)	-0.40:0.25
Collecting System	4.12 (4.21)	0.68:12.33	831.60 (465.82)	251.39:1421.86	-0.46 (0.51)	-1.10:0.19

- Rescale from disk stored value to representation unit: type = HU, intercept = -1024 , slope = 1;
- Number of bits: 12 bits (4096 values)
- Dynamic range: $-1024 \div 3071$ (float 64);
- Dimension: 512 width \times 512 height, variable number of slices;
- Voxel size: 0.35 - 0.95 mm in the the axial slice plane (usually referred as width and height), 0.65 - 1.5 mm (3.0 mm for very few data) in slice thickness (also referred as depth).

We also present some other anatomical information on aorta, cava vein, ureters and renal vessels in Table 2.3 (measurements are approximate as they are extracted from different books and studies [1, 92, 93, 110, 111, 122]).

Table 2.3: Anatomical information on aorta, cava vein, ureters and renal vessels. Measurements are approximate as they are extracted from different books and studies [1, 92, 93, 110, 111, 122].

Structure	Cross section shape	Direction major axis in coronal plane	Diameter adults [mm]	Diameter children [mm]
aorta	circular	vertical	20.6 (4.1)	11.1 (3.4)
cava vein	ellipse	vertical	19.0 (7.2) \times 13.6 (5.1)	11.6 (4.9) \times 8.2 (4.4)
ureters	circular	vertical	5.5 (1.6)	3.8 (1.0)
renal vessels	circular	horizontal	9.2 (1.9)	7.8 (1.7)

From these data we can see that the diameter of both renal vessels and ureters is very small, and given the voxel size of these images, these structures are represented in the images by a very small number of voxels compared to the entire abdominal ceCT image. Moreover, all vessels of pediatric subjects have smaller diameters than those of adults. As for the relative position, considering the axial plane of the acquired image, arteries are behind veins and the cava vein is on the left of aorta, while ureters are under arteries and veins [111]. Veins are less rigid than arteries and often compressed by tumors. This lower stiffness is also visible from the ellipsoid cross shape of the cava vein. The veins have only two bifurcations in the renal region,

for the right and left renal veins, while the arteries can also have more than two bifurcations due to possible polar renal veins [111]. The renal vessels are located approximately on the same slices and their bifurcations form approximately a 90 degree angle with the main vessels (presenting already a different principal direction than other structures). However, due to the presence of renal tumors, the main direction of the renal vessels is actually very variable.

Database availability. The PRAC Necker database is currently strictly private, and for this reason is not planned to be publicly released in the short term. In the future we plan to add segmentations done by several experts or several segmentations done by the same expert, and after appropriate approval reviews by the hospital committee, the database could be made available to the scientific community.

Chapter 3

Segmentation of kidneys and renal tumors on pediatric abdominal-visceral ceCT scanners via deep learning

The literature on the segmentation of kidneys and renal tumors on pediatric images is poor, however works on adults can be a source of inspiration. In this chapter, at first in Section 3.1 we summarize the state of the art in the segmentation of kidneys and renal tumors on adults with a focus on the deep learning challenge in the field of 2019, namely the 2019 Kidney and Kidney Tumor Segmentation Challenge (KiTS19) [53]. We identified and studied extensively the method that has stood out in the competition and became the state-of-the-art approach not only for the segmentation of these structures but also for the segmentation of medical images in general: nnU-Net [61]. Subsequently in Section 3.2 we present our implementation of nnU-Net along with a step-by-step study of the method on the KiTS19 Database. In Section 3.3 we discuss the results obtained with the winning weights of the Challenge KiTS19 on adults applied to segment the images of children through direct inference or through fine-tuning techniques. Furthermore, in Section 3.4 we propose to learn affine spatial transformations via deep learning in order to homogenize heterogeneous databases such as the children ones, in place of the use of data augmentation. Moreover, the use of an automatic bounding box detection through a proposed method allows saving time and especially memory, while keeping similar performance. Finally, in Section 3.5 we test the use of the proposed homogenization method to improve direct inference and fine-tuning on segmenting children database using the winning weights of the adults' challenge.

3.1 Related work

3.1.1 Overview

According to [55], the 3D extensions of U-Net [117] are the most used deep learning-based architectures for the segmentation of medical images, providing the best results. However, to achieve high performance with 3D CNN, large datasets are needed [126], and currently most of the pediatric datasets do not contain enough images. To overcome this limitation, transfer learning techniques from adults to children have been proposed [7, 77], but they usually require an ad-hoc and time-consuming data augmentation to take into account the

anatomical variations between children and adults. While the literature is poor on the specific problem of pediatric kidney and renal cancer segmentation, recent works on adult images are worth to be mentioned [52, 53, 95, 151]. In particular, the most recent machine learning challenges KiTS19 [53] and KiTS21 [52] address specifically the kidney and kidney tumor, intending to accelerate the progress on their automatic segmentation from ceCT images and to objectively assess the state of the art on this task.

No-newU-Net (nnU-Net) [61], a framework implementing both 2D and 3D U-Net [117], is the network that managed to obtain the best results on the challenge of 2019 [53], thanks to the use of a particular pre-processing method and an important on-the-fly data augmentation (see next section). It is worth mentioning that the second, fourth, and fifth placed methods [53] are inspired by it. Moreover, also the first three placed methods on the challenge of 2021 [52] are based on nnU-Net [61], proposing cascade versions of this with a first network for cropping and a second network for segmentation [156] or a first low resolution net to have a coarse segmentation followed by high resolution network [43, 46]. The nnU-Net was tested on 49 segmentation tasks, reaching state-of-art results in 29 of them, otherwise achieving performances on par to the top leader-board entries [61].

When working with pediatric images, the high variability in size and pose makes the distribution of data more heterogeneous compared to adult datasets. This entails a higher number of possible transformations during data augmentation, which for a network as nnU-Net (strongly based on this technique) results in a more important computational time. In addition to a difficulty in covering all the possible transformations. To tackle this problem, in [26] the authors propose to augment the convolutional kernels (instead of training data) by transforming them with several rotations. This allows the network to learn feature maps associated with different rotated versions of the input image in a single pass. However, variations in size could not be taken into account.

Before concluding, we would like to point out that we did not explore recent methods for segmentation based on Vision Transformers (ViTs) such as UNETR [49] or other networks such as ConvNeXt [88] for the very high number of parameters of their architecture (about 10 times nnU-Net). For this reason, in fact, a GPU with large VRAM (e.g. at least 32 GB) is required to be able to fit both the model and 2D input images at original size 512×512 , as well as a very large dataset to be able to reach convergence. However, in the next chapter ViT architectures are presented and tested with large adult datasets using images resized to 128×128 for the purpose of image-to-image translation [68]. Please refer to Chapter 4 for details about these methods.

3.1.2 No-new-U-Net

Since the no-new-U-Net framework [61] is the most performing network both in this precise field (kidney and renal tumor segmentation on adults) and in general on medical images, we considered it an essential part of this thesis to study it in-depth, focusing on the different techniques used by the authors to achieve state-of-art results. In fact, the authors' contribution was to automate all of the U-Net parameter choices, without needing to change the base architecture, and for this reason nnU-Net stands for "no new U-Net". So all steps from pre-processing to inference are automated, trying to apply knowledge in both deep learning and medical imaging fields to create heuristic rules to segment any type of medical images. For example the input patch size should be as large as possible while still allowing a minimum

batch size of 2 (under a given GPU memory constraint), in order to maximize the anatomical information available for decision making in the network.

While nnU-Net could be seen as a black box, we try to open this box to explain each step.

Dataset fingerprint and nnU-Net pre-processing

The idea behind nnU-Net is to provide an optimized pipeline for any medical imaging dataset by analyzing the “dataset fingerprint”, i.e. the characteristic properties of the dataset. As shown in Figure 3.1, nnU-Net is able to automatically adapt to any dataset and it takes care of all the stages of a deep learning framework. The pre-processing is the first stage and is divided into four parts. First, all the training cases are cropped to their nonzero region, which can substantially reduce the image size. Then, the algorithm records several properties of the dataset for each image, such as image size (before and after cropping), voxel size, all the classes present in the segmentation, and the intensity properties (mean, standard deviation, 99.5 percentile, and 0.5 percentile) of the “foreground” voxels (i.e. the voxels corresponding to the structures to be segmented). The next step consists in resampling the images to have the same voxel size for all the images (the physical space that a voxel represents) and then normalizing the images according to their acquisition modality. The default target voxel size for the resampling is chosen to be the median value for each axis of the voxel sizes previously recorded. For normalization of CT images, as the intensity values are quantitative and depend on the density of the tissues, nnU-Net clips the intensity values to the 0.5 and 99.5 percentiles of the foreground voxels, and normalizes them by subtracting the mean of the foreground voxels and dividing by the standard deviation of the foreground voxels (z-scoring). This particular z-scoring is done in order to have values close to or equal to zero in the structures we want to segment, facilitating the work of the network.

nnU-Net plan

The second stage is the planification of patch size, network topology and batch size, using always the dataset fingerprint previously extracted. nnU-Net prioritizes large patch sizes while remaining with a predefined GPU memory budget. The patch size is initialized as the median image size after resampling, then the first architecture is built. Downsampling in one axis is performed until further downsampling would reduce the feature map size to less than four voxels. Next, nnU-Net compares the memory footprint of the architecture built using a minimum batch size of two with a reference value. If the GPU can accommodate the batch size of two, it tests if it can accommodate larger batch sizes and it chooses the largest one that can fit into it. Otherwise, if the GPU cannot accommodate at least a batch size of two with this configuration, the patch size of the largest axis is reduced by a factor 2 and a new architecture (with one less downsampling level) is inferred with this new patch size. The process iterates until the configuration fits the predefined GPU memory budget.

nnU-Net architectures

The algorithm can train both a 3D U-Net and a 2D U-Net, and also other variants that are not analyzed here. The nnU-Net architecture closely follows the original U-Net [117] adapted to 3D images [17]. It uses two blocks per resolution step in both encoder and decoder parts, each block consisting of a convolution, instance (3D U-Net) or batch (2D) normalization, and

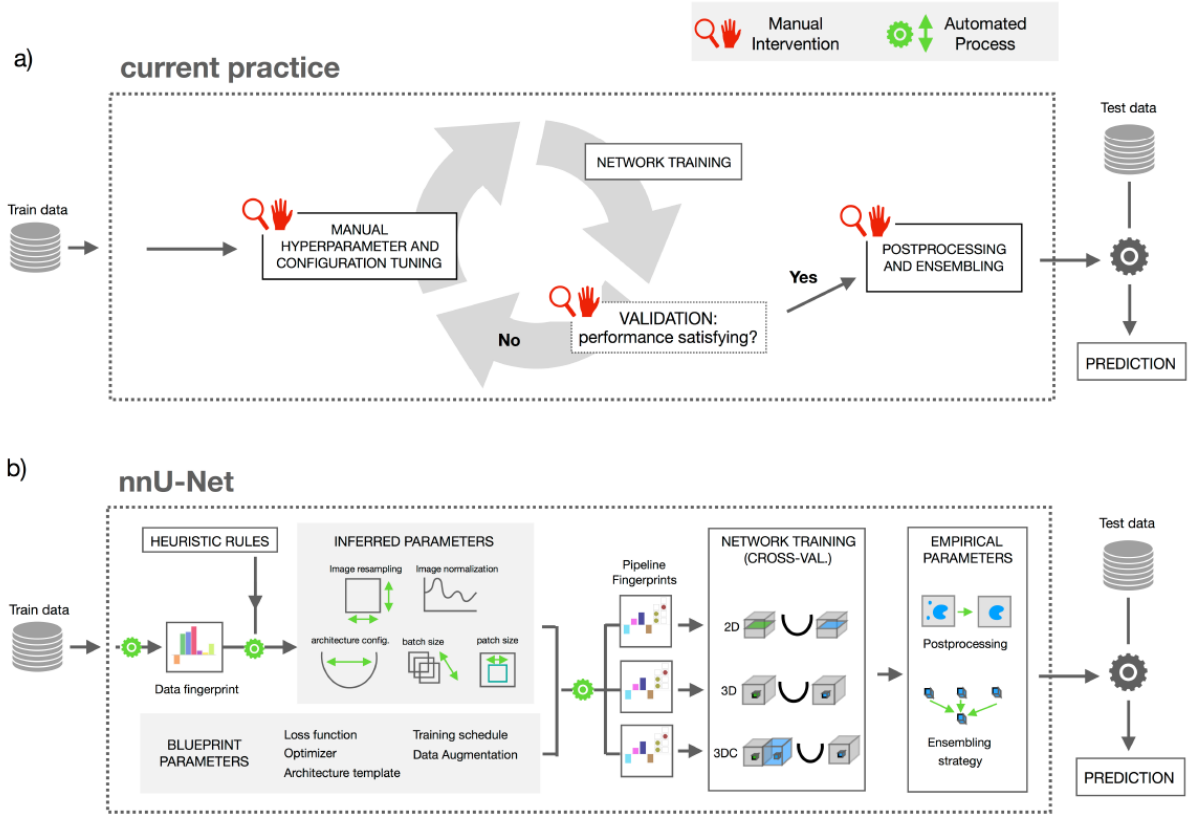


Figure 3.1: nnU-Net scheme for U-Net parameter automatic choices for: pre-processing, training, post-processing and inference. Reprinted from [61].

a leaky ReLU (negative slope of 0.01). The downsampling is done via strided convolution and the upsampling via the so-called “transposed” convolution (using the method in [154]). The initial number of feature maps is set to 32 and doubled with each downsampling step and the final number of feature maps is capped at 320 and 512, for 3D and 2D U-Net respectively, to limit the final model size. The default kernel size for convolution is $3 \times 3 (\times 3)$. The output of the network has the same number of channels as the number of classes and it is generated with a convolution operation with kernel size $1 \times 1 (\times 1)$ followed by a softmax activation (the classes are considered mutually exclusive). Training the same network to segment multiple classes instead of training multiple network to segment each class is useful for facilitating the choice in common edge voxels among multiple structures. Furthermore, this approach aligns with [69] recommendation that learning tasks with less data benefit largely from joint training on other tasks.

Besides, the network is trained with deep supervision [32], with additional outputs being added in the decoder to all but the two lowest resolutions, in order to allow the gradient to be further injected into the architecture, facilitating the training of all layers. Figure 3.2 shows two examples of architecture in 3D (on the left) and 2D (right).

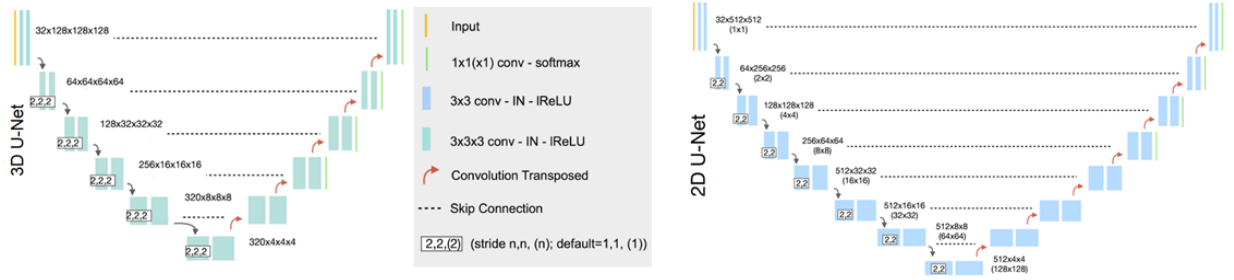


Figure 3.2: nnU-Net architectures examples for 3D (left) and 2D (right) U-Net implementations. Reprinted from [61].

nnU-Net training

The standard training schedule of nnU-Net is based on the practical experience of the authors, and includes the following components:

- **Loss function at resolution k :** $L_k = CE + SDL$ where Cross-Entropy (CE) is defined as:

$$CE = -\frac{1}{C} \sum_{i=1}^C \sum_{j=1}^M r_{ij} \cdot \log(p_{ij}) \quad (3.1)$$

and Soft Dice Loss (SDL) is derived from Dice Score (DS, see Appendix A) and defined as:

$$SDL = 1 - \frac{1}{C} \sum_{i=1}^C \frac{2 \cdot \sum_{j=1}^M |p_{ij} \cdot r_{ij}|}{\sum_{j=1}^M p_{ij}^2 + \sum_{j=1}^M |r_{ij}|} \quad (3.2)$$

where M is the number of pixels/voxels in the image, C is the number of classes, p_{ij} is the probability that a sample j is assigned to class i by the model and r_{ij} is the corresponding target sample j of class i (one-hot vector).

- **Deep supervision:** The final loss function is the sum of the loss functions computed at different levels of resolution:

$$L = \sum_{k=0}^Q w_k \cdot L_k \quad (3.3)$$

with:

$$w_k = \frac{1}{2^k} \frac{1}{\sum_{q=0}^Q \frac{1}{2^q}} \quad (3.4)$$

where $Q + 1$ is the number of resolution levels, starting from 0 as the last output level (with input image size).

- **Epochs:** 1000 with 250 mini-batches per epoch, in which each patch is randomly chosen from a random training image.

- **Optimizer:** Stochastic Gradient Descent (SGD) with a Nesterov momentum of 0.99.
- **Learning rate:** Initial learning rate lr of 0.01 reduced following the poly learning rate policy [18] decaying at each epoch e by multiplication of the initial lr by a factor of $(1 - \frac{e}{E})^{0.9}$ where E is the total number of epochs.
- **Oversampling:** To ensure that each class of interest is contained in at least 33.3% of patches in every mini-batch.
- **On-the-fly data augmentation:** This method transforms each patch in an entirely new patch that was never seen by the network before. The techniques used are applied sequentially with a certain probability of application, and include: (i) spatial data augmentation, such as rotation, scaling and mirroring; (ii) iconographic data augmentation, such as Gaussian noise and Gaussian blur, change of brightness and contrast, simulation of low resolution images and gamma augmentation. The data augmentation is implemented with the *batchgenerators* framework [62]. Details are available in the Appendix B.

During inference, nnU-Net uses a sliding window to extract patches with overlapping of half the size of the patch. Moreover, to reduce artifacts and reduce the influence of position of patches close to the image borders, a Gaussian importance weighting is applied to the window, increasing the weight of the center voxels in the softmax aggregation. Test-Time Augmentation (TTA) [137] is also applied by mirroring along all axes. This technique helps to reduce errors by presenting the same image under different points of view to the network, especially in 3D when patches are used and spatial information is partially lost. Furthermore, this technique is also useful in 2D when an instance to be segmented can be found in different parts of the image, e.g. unilateral renal tumor of a certain length and size.

3.2 Application of no-new-U-Net on a database of adult images

After understanding the details of the nnU-Net we decided to reproduce it from scratch, and apply it to the KiTS19 challenge database, available online (<https://github.com/neheller/kits19>), both to be able to understand through a step-by-study which parts really play a fundamental role in the method, and to have a first approach to the segmentation of kidneys and renal tumors.

3.2.1 Adults database of KiTS19 Challenge

We used the dataset that was built for the KiTS19 challenge for comparing our reproduction of nnU-Net with the results obtained using the implementation available on GitHub (<https://github.com/MIC-DKFZ/nnU-Net>). The “grand challenge MICCAI” KiTS (Kidney and Tumor Segmentation) Database of 2019 [53] is composed of abdominal ceCT scanners (with late arterial/arteriovenous phase imaging, see details in Section 1) of 300 patients (180 males and 120 females, aged between 50 and 70 years old) with renal tumors. Images have been acquired just before surgery (RN or NSS) in more than 50 referring institutions. A reference manual segmentation for kidneys and renal tumors is available for 210 patients. The

characteristics of the images are very diverse in terms of voxel dimensions, contrast timing and scanner field of view. The voxel size has a wide range: 0.5 - 5.0 mm in slice thickness and 0.65 - 0.95 mm in the axial slice plane, with a size of 512×512 pixels and variable number of slices. For what concerns contrast intensity, these images present slightly lower HU values for both structures but falling within the standard deviation of children's database. Moreover, although these images were acquired with a higher radiation dose, thus achieving higher quality, compared with that of children there is no severe difference due to the small body size of the pediatric patients. The major differences with children images lie in tumor size and renal parenchyma deformation. Examples of these images are visible in Figure 3.3.

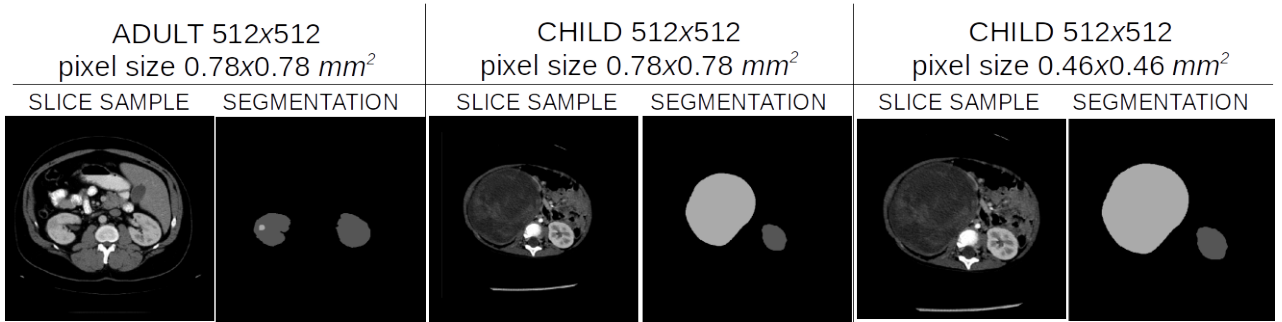


Figure 3.3: Examples of adults and children images using different pixel sizes. In the segmentation image, kidneys are in dark grey while renal tumor in light grey.

3.2.2 Results with the KiTS19 dataset from using and rebuilding nnU-Net

First of all, since the results available in the challenge [53] are on the 70 test data that are not downloadable and the weights made available on GitHub are trained on all the 210 images with reference segmentation, we decided to redo the experiences using the 210 labeled images for training, validation and test sets. In particular we have divided the dataset in 162 for training, 18 for validation and 30 for test set.

Table 3.1 shows the results obtained with batch size of 4, comparing the original framework with our step-by-step reproduction, both in 2D and 3D. The Dice score is used as evaluation criterion.

We can see that ensuring oversampling is critical to have a model that performs well, mainly on tumors, otherwise there are not enough images with tumors. It is important to say that this oversampling method necessarily requires the use of on-the-fly data augmentation since otherwise there would be a high risk of overfitting. The addition of the deep supervision to prevent vanishing gradient, and subsequently of the instance normalization to make the outliers (very present in 3D patch training) less impactful, increases by a few percentage points both Dice scores. Using also bias and changing the lr with the poly lr policy instead of a fix lr significantly improves the results. Finally, inference with test-time augmentation (TTA) techniques gives the best results, that are comparable to the original implementation for both 2D and 3D U-Net, with slight differences due to the random selection of patches and parameters of data augmentation through epochs. Other experiments have also demonstrated that the use of the CE as regularizer speeds up convergence of the training, making it more stable, while without the use of the data augmentation there is a decrease in performance of

1-2% for the kidney and 4-5% for the tumor; moreover the TTA cannot be used if the network is not trained even with mirrored images.

Table 3.1: Dice score (DS) in % (mean and standard deviation) obtained for the 30 patients of the test set after training the 2D and the 3D full resolution U-Net of nnU-Net, using both the out-of-box nnU-Net and our adaptation. According to [61], we have gradually added the techniques in order of decreasing importance.

Network	Oversampling	Deep Sup.	Normalization	Bias	Poly lr	TTA	DS Kidney [mean (sd)]	DS Tumor [mean (sd)]
Original 3D	x	x	Instance	x	x	x	95.59 (5.34)	72.12 (23.86)
Our 3D			Batch				84.45(8.76)	36.78(35.04)
Our 3D	x		Batch				90.49(6.62)	55.79(37.09)
Our 3D	x	x	Batch				91.89(3.04)	55.73(37.42)
Our 3D	x	x	Instance				91.76(2.30)	56.37(32.84)
Our 3D	x	x	Instance	x			93.49 (8.28)	61.96 (29.76)
Our 3D	x	x	Instance	x	x		93.71 (6.78)	66.16 (24.99)
Our 3D	x	x	Instance	x	x	x	94.86 (5.71)	68.54 (27.36)
Original 2D	x	x	Batch	x	x	x	92.77 (17.42)	67.81 (27.92)
Our 2D			Batch				92.42(4.94)	36.84(36.87)
Our 2D	x	x	Batch	x	x	x	94.54 (3.87)	65.21 (31.99)

3.3 Transfer learning from adults to children

We now address the first main objective of the thesis, i.e. the segmentation of kidneys and renal tumors in pediatric CT images. Our first approach is based on transfer learning, from adults to children.

In our first experiments, we used the 3D no-newU-Net [61] trained on adults, with the weights winner of the KiTS19 challenge [53], directly on the children images (same weights), and then using transfer learning (fine tuning of the weights [44]). In this case we put ourselves in the position where we cannot have access to the training data but only to the weights of the pre-trained network. This case reflects more the reality where healthcare centers do not easily allow data exchange due to the private and sensitive nature of data and the still low security in data transfer and communication systems [123]. Research is advancing in this direction for example with the so-called federated learning [123]. Nevertheless, for the sake of completeness, we also tried a training using all the available images (i.e. adults and children database together) that brought less performing results than all those that are shown in this section, because of the difference between the two domains. Due to the above reason, methods for reducing distance between the two domains, that require the availability of both databases, such as those presented in cite [40, 130], have not been explored.

We used a learning rate lr of 0.01 as in the training (lower lr resulted in lower performance). The weights were obtained using as pre-processing a common voxel size of $3.22 \times 1.62 \times 1.62 \text{ mm}^3$ (depth \times width \times height), a clipping in range $[-78, 303]$, and a “foreground” voxel mean of 99.9 and a standard deviation 77.71 for z-scoring normalization. A patch size of $80 \times 160 \times 160$ was used. We divided the set of 80 children images presented in Chapter 2 into training (58), validation (7) and test set (15), according to tumor location and patient’s age (different sizes of organs and abdomen), in order to have a balanced division of the dataset as shown in Table 3.2.

The results, evaluated using the Dice score in Table 3.3, show that only when we fine-tune most of the weights the results become good, but still not good enough for clinicians. This confirms the important differences between adults and children images, as shown in Figures 1.2

Table 3.2: Division in groups of the children database for balanced datasets. Rows are divided according to patient’s age and columns according to tumor location. For each cell: total number, and in brackets the distribution for training, validation and test sets.

patient age	right tumor	left tumor	bilateral tumor
less than 2 y.o.	17(13,1,3)	17(13,1,3)	5(3,1,1)
between 2 and 5 y.o.	17(13,1,3)	7(5,1,1)	0
more than 5 y.o.	8(5,1,2)	9(6,1,2)	0

and 3.3, and that it is not always possible to use transfer learning strategies to overcome the problems of limited and heterogeneous data.

Table 3.3: Results (mean and standard deviation of Dice score) using weights of 3D nnU-Net trained on adults KiTS database [53]. More results can be found in the Appendix C.

Technique (3D Networks)	Dice Score Kidney	Dice Score Tumor
Direct Inference (weights frozen)	20.83 (35.55)	18.29 (35.73)
Fine-Tuning (first 2 encoder blocks and last decoder 2)	53.38 (25.84)	51.05 (31.76)
Fine-Tuning (entire decoder)	81.75 (7.18)	75.79 (23.24)
Fine-Tuning (entire network)	84.99 (6.38)	81.08 (23.01)

As we mentioned, non-satisfactory results using direct inference or simple technical transfer learning are due to evident differences in tumor size between adults and children. In the former the tumor appears much smaller but we must keep in mind that all the other organs are larger. These tumors have a diameter between 2.6 and 6.1 *cm*, which is comparable to the size of a kidney in children from a few days of life to 6 years old. Moreover, as shown in Figure 3.4 (left group) if we compare images of the same dimension, for example 160×160 , and homogenizing them to the same pixel size, $1.62 \times 1.62 \text{ mm}^2$, they appear totally different visually, emphasizing the difference also in the size of the abdomen, and not only between children and adults but also between children of different ages. To have a patch in a child image of the same dimension (160×160) that appears similar to that of adults, we should use a smaller pixel size, for example $1.0 \times 1.0 \text{ mm}^2$, as illustrated in Figure 3.4 on the right. To clarify this point it is worth remembering that during the pre-processing step the images are resampled to a common voxel size.

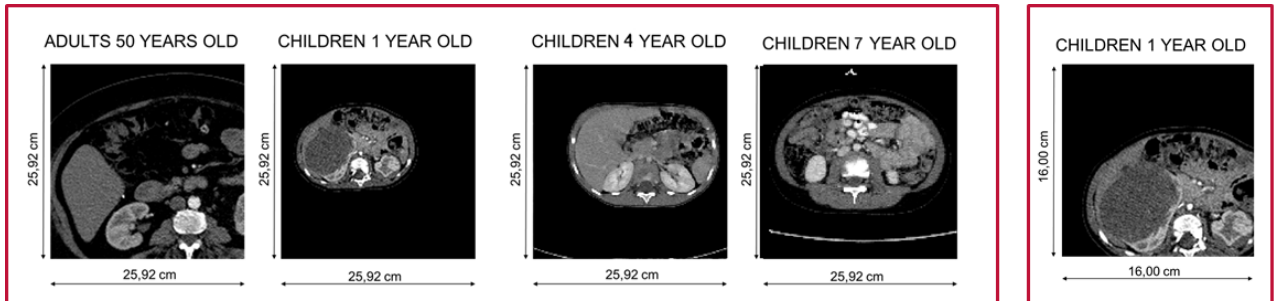


Figure 3.4: Left box: comparison of 160×160 images with a pixel size of $1.62 \times 1.62 \text{ mm}^2$. Right box: a 160×160 child image with a pixel size of $1.0 \times 1.0 \text{ mm}^2$.

As further illustration, we did some other tests (Table 3.4) with direct inference and fine-tuning of the entire network but in 2D (512×512) this time and using a smaller pixel size

of $0.46 \times 0.46 \text{ mm}^2$ for children images, while the weights used as baseline are from KiTS19 2D nnU-Net on adults with pixel size of $0.78 \times 0.78 \text{ mm}^2$. In order to check for contrast differences as well, we also did some tests using an intensity normalization calculated on the pediatric training-set. This resulted in clipping to the range $[-36, 303]$, and a z-scoring using a “foreground” voxel mean of 76.99 and a standard deviation of 67.73. The idea behind these other tests is that the intensity in kidney and tumors is different in children and therefore a different normalization is needed to bring the values of “foreground” voxels similar to those in adults.

Table 3.4: Results (mean and standard deviation of Dice score) on children test images with same pixel size and intensity normalization as adults and with smaller pixel size of $0.46 \times 0.46 \text{ mm}^2$ and different normalization, using weights of 2D nnU-Net trained on adults KiTS19 database [53] with pixel size of $0.78 \times 0.78 \text{ mm}^2$.

Technique (2D Networks)	Children pixel size (mm^2)	Children intensity normalization	Dice Score Kidney	Dice Score Tumor
Direct Inference (weights frozen)	0.78×0.78	as adult dataset	37.43 (29.95)	36.88 (24.91)
Direct Inference (weights frozen)	0.46×0.46	as adult dataset	48.51 (26.33)	42.75 (33.01)
Direct Inference (weights frozen)	0.46×0.46	as children dataset	41.00 (29.20)	39.24 (31.84)
Fine-Tuning (entire network)	0.78×0.78	as adult dataset	89.62 (3.92)	78.72 (18.72)
Fine-Tuning (entire network)	0.46×0.46	as adult dataset	90.01 (3.57)	79.89 (21.77)
Fine-Tuning (entire network)	0.46×0.46	as children dataset	86.90 (5.22)	70.00 (29.13)

Results show that the use of a smaller pixel size leads to better results, while using a different normalization leads to worse results, confirming the previous assumptions.

We also want to emphasize that all transfer learning results are worse than the results without using it, i.e. training the network directly with the Necker PRAC dataset. Figure 3.5 shows a study done on Necker PRAC database with nnU-Net by increasing the number of patients used during training from time to time. Using all available patients (65) the results on Dice score in 2D are of 89.59% (with a standard deviation of 4.21%) for kidneys and 82.49% (19.02%) for tumors, while in 3D the results are of 90.15% (3.57%) for kidneys and 86.92% (10.49%) for tumors (for comparison see previous Tables 3.3 and 3.4). Several considerations can be gleaned from this study. First, 3D network shows better performance compared to 2D only when an high number of training patients is available, particularly with regard to lower variability of results. However, for tumor segmentation the standard deviation remains very high, stating a strong variability of the results. A second consideration is that as the number of images increases, the performance improves for both 2D and 3D networks, particularly on tumor segmentation results. This means that efficient transfer learning could be useful in cases when the number of patients is limited, as is usually the case in smaller hospitals or research centers than Necker Hospital.

From all these results three conclusions can be drawn:

1. spatial transformations are necessary to make transfer learning between adults and children possible, and this idea is examined in Section 3.5;
2. by contrast, there are no significant differences in image contrast, and since adults values used for normalization are extracted from a larger dataset they are more suitable also for transfer learning on children images;
3. as shown in Figure 3.4, the heterogeneity in pose and size of the pediatric images makes

it difficult to segment them, in particular for some renal cancers, even through the use of data augmentation techniques; Section 3.4 presents our study of this point.

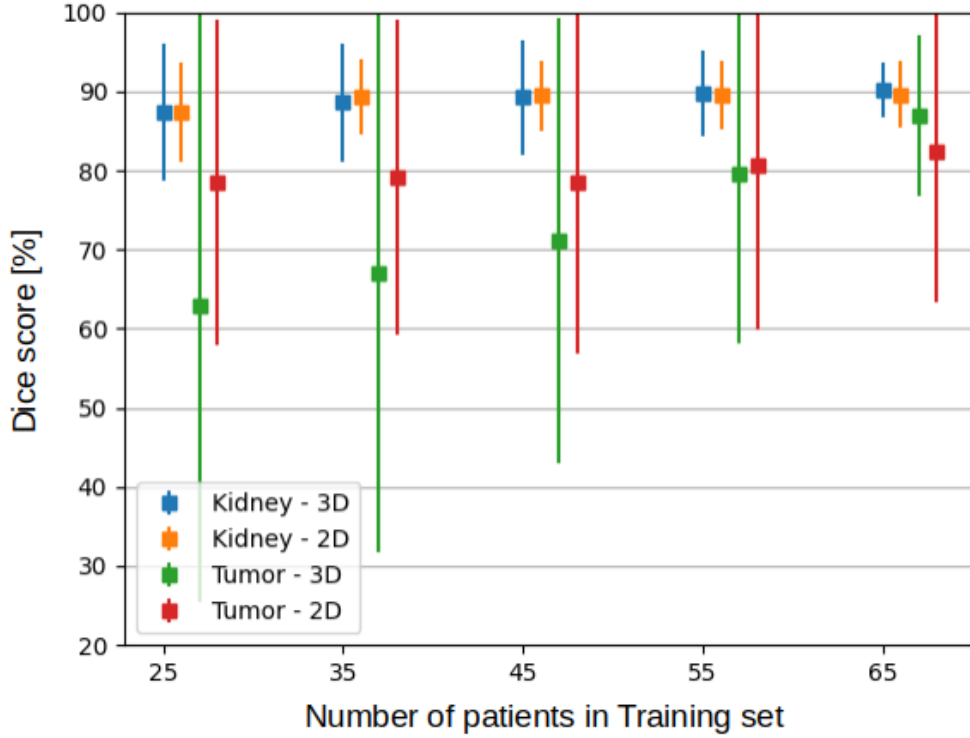


Figure 3.5: Dice scores (y -axis in mean and standard deviation) in a study done on the Necker database with nnU-Net [61] by increasing the number of patients used during training from time to time (x -axis).

3.4 Automatic size and pose homogenization with Spatial Transformer Network to improve and accelerate pediatric image segmentation

For all the reasons presented in the previous sections, we propose to take a different perspective with respect to the usual data augmentation strategy. Instead of *augmenting* the number of training images to cover the entire data distribution, we propose to *reduce* the data variability through an homogenization in terms of size and pose. In this way, unlike the original nnU-Net, which from a finite training dataset strives to learn a complex mapping, including pose and scale variations (translation is an inherited property of convolutions), our segmentation module learns a simpler mapping focusing on images with normalized pose and size. Moreover, the U-Net is capable of reasoning between different objects in its receptive field, using absolute position and directional relationships to ensure proper segmentation [64, 113, 116]. The homogenization in pose and scale can facilitate this behavior. In order to do that, we first *learn* an optimal similarity transformation (without reflection) to a clinically relevant reference subject. Then, to accelerate the segmentation, we also *learn* to crop the region of interest (ROI) as a square patch which is used as input image for the final segmentation network instead of the original (bigger) image. We propose a new architecture composed of

three neural networks: a first Spatial Transformer Network (STN) [65] that deals with homogenization of pose and size; a second STN that crops the homogenized image in the region of interest (ROI); and finally a segmentation network, built as a nnU-Net [61], in which the cropped homogenized image is given as input and the output is then restored to its original pose and size, and uncropped, using the inverse of the two transformation matrices previously computed. This original combination allows us to deal with small and heterogenous datasets, and showed some interesting results. This approach led to a paper accepted at the IEEE International Symposium of Biomedical Imaging (ISBI) 2021 [ISBI-21].

3.4.1 Proposed method

Pre-processing

All images are pre-processed as for the nnU-Net [61] previously presented, i.e. (i) a non-zero region cropping, (ii) a resampling of the images to have the same pixel size, (iii) a clipping of the intensity values to the 0.5 and 99.5 percentiles of the foreground voxels, and (iv) a Z-scoring normalization.

Architecture

The proposed framework is presented in Figure 3.6. We now present the three networks in detail.

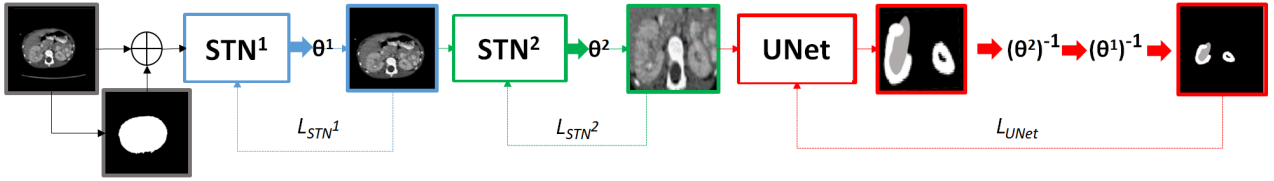


Figure 3.6: Schema of our proposed framework.

STN to homogenize pose and size At first a Spatial Transformer Network (STN) deals with homogenization, transforming all images to be as similar as possible in size and pose to a chosen one (STN¹ in Figure 3.6). The reference image was chosen among patients aged 2 years, who represent the average in the database, and among them a patient with the best pose was chosen, according to the medical doctors' directives. This STN is composed of a localization network, composed of an encoder with two stacked convolutional blocks with MaxPooling and ReLU, which reduces the image by a factor of 4, and two fully convolutional layers, as in [65]. This network regresses five values (1 value \mathcal{R}_{xy} for angle, \mathcal{S}_x and \mathcal{S}_y for scaling, and \mathcal{T}_x and \mathcal{T}_y for translation) in 2D and nine in 3D (other two values \mathcal{R}_{yz} and \mathcal{R}_{zx} for angles, one \mathcal{S}_z for scaling and one \mathcal{T}_z for translation), defining the transformation matrix θ^1 . Then, we proceed as done in the original STN [65]. First a grid generator iterates over a regular grid (normalized in range $[-1,1]$, named sampling grid) of the output image and uses the inverse transformation θ^{-1} to calculate the corresponding sample (normalized) positions in the input image. Please note that the size of the sampling grid determines the size of the target image,

that, thanks to the use of normalized coordinates, does not necessarily need to be the same size as the one of the input image. Subsequently, a differentiable image sampling is performed: a sampler iterates over the entries of the sampling grid and extracts the corresponding voxel values from the input map using a chosen interpolation method (usually bilinear for images and nearest-neighbor for binary masks) to create the final output image. The transformation matrix θ^1 is defined in 2D as:

$$\begin{bmatrix} \mathcal{S}_x \cos \mathcal{R}_{xy} & -\mathcal{S}_x \sin \mathcal{R}_{xy} & \mathcal{T}_x \\ \mathcal{S}_y \sin \mathcal{R}_{xy} & \mathcal{S}_y \cos \mathcal{R}_{xy} & \mathcal{T}_y \\ 0 & 0 & 1 \end{bmatrix} \quad (3.5)$$

and in 3D as:

$$\begin{bmatrix} \mathcal{S}_x(\cos \mathcal{R}_{xy} \cos \mathcal{R}_{yz}) & \mathcal{S}_x(\cos \mathcal{R}_{xy} \sin \mathcal{R}_{yz} \sin \mathcal{R}_{zx} - \sin \mathcal{R}_{xy} \cos \mathcal{R}_{zx}) & \mathcal{S}_x(\cos \mathcal{R}_{xy} \sin \mathcal{R}_{yz} \cos \mathcal{R}_{zx} + \sin \mathcal{R}_{xy} \sin \mathcal{R}_{zx}) & \mathcal{T}_x \\ \mathcal{S}_y(\sin \mathcal{R}_{xy} \cos \mathcal{R}_{yz}) & \mathcal{S}_y(\sin \mathcal{R}_{xy} \sin \mathcal{R}_{yz} \sin \mathcal{R}_{zx} + \cos \mathcal{R}_{xy} \cos \mathcal{R}_{zx}) & \mathcal{S}_y(\sin \mathcal{R}_{xy} \sin \mathcal{R}_{yz} \cos \mathcal{R}_{zx} - \cos \mathcal{R}_{xy} \sin \mathcal{R}_{zx}) & \mathcal{T}_y \\ -\mathcal{S}_z \sin \mathcal{R}_{yz} & \mathcal{S}_z(\sin \mathcal{R}_{xy} \cos \mathcal{R}_{yz}) & \mathcal{S}_z(\cos \mathcal{R}_{xy} \cos \mathcal{R}_{yz}) & \mathcal{T}_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (3.6)$$

The input of the STN is composed of the original image concatenated with its “foreground mask”, a binary mask representing the abdomen and easily computed as the largest connected component of a thresholded image. The network is optimized using a Soft Dice loss (see Equation 3.2) function L_{STN^1} between the homogenized output “foreground mask” and the “foreground mask” of the reference image:

$$L_{STN^1} = SDL(\theta^1 I_{fm}, T_{fm}) = SDL(\mathcal{H}_{fm}, T_{fm}) \quad (3.7)$$

where I_{fm} is the input “foreground mask”, θ^1 is the predicted matrix, $\mathcal{H}_{fm} = \theta^1 I_{fm}$ is the homogenized output “foreground mask”, and T_{fm} is the reference “foreground mask”. The method is illustrated in Figure 3.7.

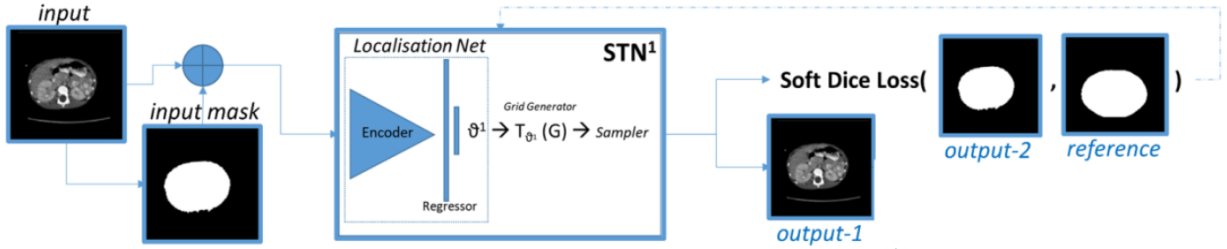


Figure 3.7: Detailed schema of our STN^1 framework.

STN for ROI cropping Then, a second STN crops the homogenized image in the region of interest (ROI), where the structures to be segmented are present (STN² in Figure 3.6). This network is the same as the previous one but it regresses 4 values for 2D (6 for 3D): 2 (resp. 3) for scaling and 2 (resp. 3) for translation, that are used to construct a scaling and translation matrix θ^2 for cropping. A target matrix and the associated target bounding box are automatically calculated using the minimum and maximum non-zero values of the reference segmentation. However, the minimum crop size is considered to be a quarter of the original image. This allows not deforming the image too much; in fact, as explained in the

previous paragraph, the size of the sampling grid (and therefore of the output images) must be previously chosen and must be the same for all images. This can result in a final image with a difference ratio between the axes. We underline that in our method the user can choose whether to keep the image in its original size, halve it or reduce it to a quarter (minimum size of the patches coming out of the STN). In the first case this allows localization and zooming on target structures facilitating the segmentation task, while in the other cases this permits reducing time and memory requested for the segmentation network. In fact, object recognition methods such as FasterRCNN [114] or nnDetection [6] only allow bounding box recognition; furthermore their use in concatenation with a segmentation network is not trivial. To the best of our knowledge there is no method similar to our proposition in the literature.

The values of θ^2 are computed from to the vertices of the bounding box, in 2D case as:

$$\begin{bmatrix} \frac{x_{max}-x_{min}}{\mathcal{W}} & 0 & \frac{x_{max}+x_{min}}{\mathcal{W}} - 1 \\ 0 & \frac{y_{max}-y_{min}}{\mathcal{H}} & \frac{y_{max}+y_{min}}{\mathcal{H}} - 1 \\ 0 & 0 & 1 \end{bmatrix} \quad (3.8)$$

and in 3D as:

$$\begin{bmatrix} \frac{x_{max}-x_{min}}{\mathcal{W}} & 0 & 0 & \frac{x_{max}+x_{min}}{\mathcal{W}} - 1 \\ 0 & \frac{y_{max}-y_{min}}{\mathcal{H}} & 0 & \frac{y_{max}+y_{min}}{\mathcal{H}} - 1 \\ 0 & 0 & \frac{z_{max}-z_{min}}{\mathcal{D}} & \frac{z_{max}+z_{min}}{\mathcal{D}} - 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (3.9)$$

where x_{min} , y_{min} and z_{min} are the minimal coordinates of the bounding box for all directions and x_{max} , y_{max} and z_{max} are the maximal coordinates, while \mathcal{W} , \mathcal{H} and \mathcal{D} are the image sizes. This second STN for the cropping is trained using the loss function L_{STN^2} , defined as the sum of a L^1 term (mean absolute error) between the cropped output image and the target crop, and a L^2 term (root mean squared error) between the “scaling and translation” output matrix and the “scaling and translation” target matrix:

$$L_{STN^2} = \frac{1}{N} \sum_{n=1}^N \|\mathcal{H}_{C_n} - T_{C_n}\|_2 + w_C \cdot \sqrt{\frac{1}{N} \sum_{n=1}^N \left\| \left(\theta_n^2 - \theta_n^T \right) \right\|_2^2} \quad (3.10)$$

where θ^2 is the predicted matrix, \mathcal{H}_C is the cropped output, θ^T is the target matrix, T_C is the target crop and N is the batch size. The L^2 term is weighted by a factor w_C . This combination was proved experimentally efficient, probably due to the robustness of the L^1 norm to outliers. The method is illustrated in Figure 3.8.

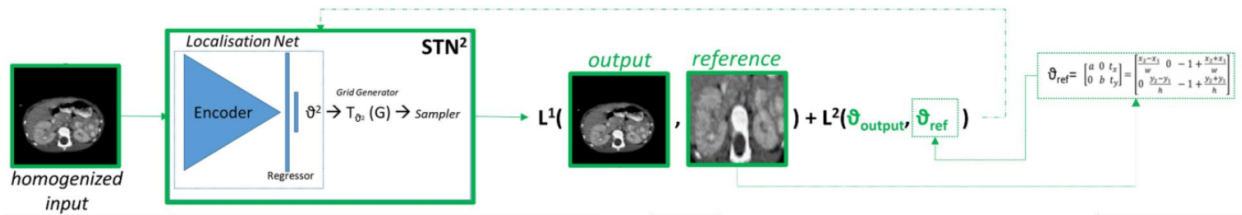


Figure 3.8: Detailed schema of our STN^2 framework.

U-Net for segmentation At the end of our framework a U-Net takes as input for the segmentation the cropped homogenized image and the output is then restored to its original pose and size, and uncropped, using the inverse of the two transformation matrices previously calculated. In this way, the predicted segmentation is directly compared with the original reference segmentation, without the need of transforming this and thus losing information. The U-Net is constructed using the tool “planes” of nnU-Net, as described in Section 3.1.2, and it is optimized as in the nnU-Net training using a loss function L_{U-Net} defined as the sum of cross entropy CE and Soft Dice, both between prediction and reference segmentation:

$$L_{U-Net} = \sum_{q=0}^{Q-1} \frac{1}{2^q} \left(CE(\theta^{1-1}(\theta^{2-1}P_q), R) + SDL(\theta^{1-1}(\theta^{2-1}P_q), R) \right) \quad (3.11)$$

where R is the reference segmentation, P_q is the prediction at the q level of resolution (considering the output layer of the network as 0 level). We use the Deep Supervision technique [32] up to the level $Q - 1$, where there is the last skip connection.

Training

For training the STNs, the best solution has been experimentally identified as a training of one STN after the other for 50 epochs using a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01. For STN^2 the best w_C was empirically found as 10. The U-Net is trained as detailed in Section 3.1.2 with an early-stopping condition, unlike the original network since on-the-fly data augmentation is not used. In addition, at each epoch all available images are examined, but the total number of iterations of the original nnU-Net is left unchanged for all trainings. This is done in order not to make the trainings dependent on the patches randomly selected at each epoch. Moreover, all the seeds in the random number generator were fixed for reproducibility. We used a batch size of 12 with 2D slices of size 512×512 and 2 with 3D images of resized at $128 \times 128 \times 128$ (further details in the next section). We used the Necker PRAC database divided as presented in Section 3.3 with 58 subjects for training, 7 for validation and 15 for test. For 2D experiments these resulted in 15036, 3760 and 5310 slices, respectively.

All trainings and tests were done using PyTorch framework [108] and run on Télécom server clusters which have a GPU NVIDIA® Tesla® P100 with 16 GB of VRAM and a CPU Intel(R)® Xeon(R)® E5-2643 v4 @ 3.40GHz with 6 cores and 126 GB of RAM. These technical specifications on framework, GPU and CPU are the same for the other chapters of this thesis.

3.4.2 Results and discussion

STN_{pose-size} We first tested the quality of the pose-size homogenization done with the proposed STN^1 method and we compared it with a *non-deep learning* state-of-the-art medical image registration algorithm named Simple-Elastix [96]. We tested this using both Dice score and Mattes Mutual Information [97] (MMI) as cost function (more parameters details are available in Appendix B). Please note that other deep learning methods such as VoxelMorph [4] have not been tested since they also use deformable transformation, such as elastic warping which are destructive and cannot be inverted in some cases [65]. The images of the test set were used at the original size of 512×512 for 2D tests, while in 3D the depth dimension was adjusted (removing ending slices or adding empty slices) to have an image of size $512 \times 512 \times 512$. This

allows us to resize the image to a size of $128 \times 128 \times 128$ without deforming it and train the STN^1 with these images, saving time and memory. In fact, as explained in Section 3.4.1, both the sampling grid and input positions are normalized between -1 and 1 . This allows the parameters of θ^1 to be calculated on the smaller similar image of size $128 \times 128 \times 128$, and then we apply the steps of grid generator and sampling using those parameters to the original $512 \times 512 \times 512$ image in validation and inference. We believe that the loss of information is neglectable for the homogenization objective of pose and size of the whole body, as well as for cropping.

In Table 3.5 results are shown as mean and standard deviation of the Dice score, mean square error (MSE) and time per subject. In case of 2D images, the latter parameter refers to the homogenization of the total number of slices present in a patient (pre-processing is also included).

Table 3.5: Results (mean and standard deviation of Dice score, mean square error (MSE) and test time per subject) on the Necker PRAC test set of 15 subjects using the proposed STN^1 and Simple-Elastix [96] methods for pose-size homogenization. MMI is the Mattes Mutual Information [97].

2D images				
Method	Cost function	Dice score [%]	MSE	Time per subject
Simple-Elastix [96]	MMI	88.10 (3.30)	0.08 (0.01)	5 m 45 s (3 m)
Simple-Elastix [96]	Dice	90.10 (2.21)	0.07 (0.01)	3 m 35 s (2 m)
STN pose-size	Soft Dice	94.55 (4.72)	0.04 (0.03)	34 s (16 s)
3D images				
Method	Cost function	Dice score [%]	MSE	Time per subject
Simple-Elastix [96]	MMI	67.07 (10.89)	0.10 (0.03)	2 m 32 s (46 s)
Simple-Elastix [96]	Dice	68.39 (10.19)	0.09 (0.03)	3 m 2 s (1 m)
STN pose-size	Soft Dice	71.12 (10.88)	0.08 (0.04)	48 s (14 s)

Our method shows both better registration performance between “foreground masks” and a significantly shorter time to register all the slices of a subject. The pre-processing time lasts about 26 seconds, so the inference time of our network is less than 10 seconds for 2D images and about 12 seconds for 3D images. This allows our technique to be used in conjunction with a segmentation network during both training and inference, without altering too much the time performance of this. 3D performances are lower because of the greater variability in body length as well as the limited number of volumetric images available to train our network. The training time for such a network is only 3 hours in 2D and 5 hours in 3D, to be done only once. Moreover, in case we want to train a segmentation network with homogenized images, as we present later in this section, we can use the network in inference as shown in Figure 3.9. This avoids to previously register all the 2D images which would require about 15h as well as an additional storage memory of 63.4 GB (or 8.8 GB if compressed images) in order to preserve the original images. Some results are available in Appendix C.

STNcrop We then tested the proposed STN^2 for ROI cropping and we compared as backbone network and training method our STN , inspired from [65], with FasterRCNN [114] for 2D slices and nnDetection [6] for 3D images, the state-of-the-art deep learning methods for such a task. The methods are also tested using pre-trained STN^1 in order to homogenize the images and facilitate ROI detection. It is important to note that both FasterRCNN and

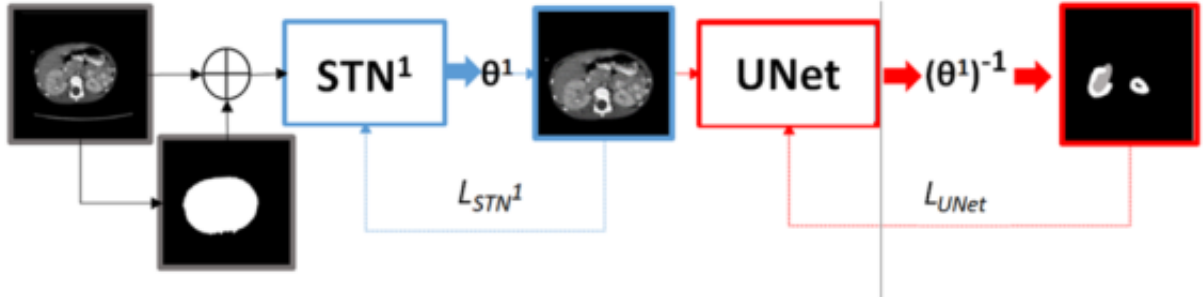


Figure 3.9: Schema of our $STN^1 + U\text{-Net}$ framework.

nnDetection are used only to calculate the minimum and maximum values of the bounding box and thus the θ^2 matrix presented in Equations 3.8 and 3.9. For 2D tests, images were used at the original size of 512×512 , while for 3D they were carried out as explained in the previous paragraph. In Table 3.6 results are shown as mean and standard deviation of L1 error for images, MSE for θ^2 , and time per subject. As for the previous study, the latter parameter refers to total number of 2D slices present in a patient (and pre-processing is included).

Table 3.6: Results (mean and standard deviation of L1 error, mean square error (MSE) and test time per subject) on Necker PRAC test set of 15 subjects using the proposed STN^2 , FasterRCNN [114] and nnDetection[6] methods for ROI cropping.

Method	Backbone	L1 (image)	MSE (θ)	Time per subject
2D images				
STN crop	FasterRCNN [114]	0.42 (0.58)	0.03 (0.09)	44 s (18 s)
STN pose-size + STN crop	FasterRCNN [114]	0.32 (0.47)	0.02 (0.07)	46 s (18 s)
STN crop	STN [65]	0.64 (0.39)	0.02 (0.04)	30 s (13 s)
STN pose-size + STN crop	STN [65]	0.54 (0.36)	0.02 (0.03)	32 s (13 s)
3D images				
STN crop	nnDetection [6]	1.51 (0.44)	0.12 (0.05)	55 s (13 s)
STN pose-size + STN crop	nnDetection [6]	1.34 (0.25)	0.11 (0.06)	1 m 5 s (14 s)
STN crop	STN [65]	1.22 (0.19)	0.06 (0.03)	41 s (13 s)
STN pose-size + STN crop	STN [65]	1.14 (0.22)	0.05 (0.04)	51 s (14 s)

Given its more particular architecture (authors [114] use so-called ROI Pooling layers that extracts equal-length feature vectors from proposed ROIs extracted from the input image, which result also in more parameters), FasterRCNN performs better than the classical STN architecture in finding the minimum and maximum values of the bounding box in 2D images. Such a network succeeds in individuating structures even if only a few pieces are present in the image, however, this sometimes leads to mistakenly considering certain structures as kidneys or tumors. This error can be seen in the high standard deviation in the third and fourth columns of Table 3.6, as well as a higher mean MSE on the θ^2 matrix. The use of STN^1 as first step significantly improves the results for both 2D architectures, with only a few more seconds per subject (considering also here about 26 seconds for pre-processing). The training time for the STN backbone network is only 3 hours, while it is 9 hours for FasterRCNN. For all the 2D segmentation tests shown in the next paragraph, FasterRCNN was chosen as backbone network for STN^2 .

Focusing now on 3D results, the number of volumetric images proved to be not sufficient

to obtain satisfactory results, particularly for nnDetection, which works similarly to FasterRCNN and thus has much more parameters than the original STN backbone. The graphs in Figure 3.10 show that we fall into overfitting problems for both methods. These difficulties may also be due to resampling to have image size of $128 \times 128 \times 128$, but limitations in memory do not allow training using the images with their original size of $512 \times 512 \times 512$. In such a case, the use of STN^1 as a first step does not lead to significant improvements, given also the not optimal performances of STN^1 network in 3D. For all these reasons, the use of the proposed STN methods to improve segmentation are not explored in the 3D scenario. However, the use of STN^1 and STN^2 in 3D is used in the 3DSlicer [37] plug-in developed for the IMAG2 lab of Necker hospital as an initialization for the bounding box (later adjusted by the user) to select the ROI in which to extract the 3D patches. More details are provided in Chapter 6. Some results are available in Appendix C.

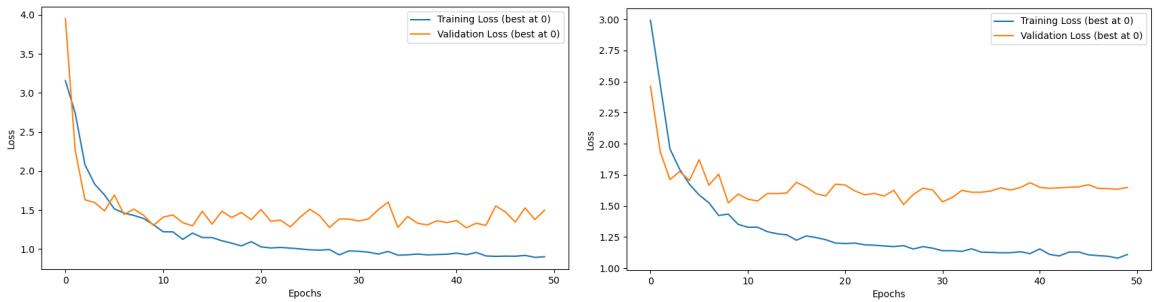


Figure 3.10: Train and validation loss values for STN^2 (left) and nnDetection [6] (right) for the 3D scenario.

Segmentation The first set of segmentation experiments was to test the size and pose homogenization STN^1 network on the Necker PRAC database. For the reason explained in the previous paragraph, we tested only 2D networks. Images were used at the original size (512×512). The results are shown in Table 3.7 in Dice score, precision, recall and the 95th percentile of Hausdorff distance (95HD). Further details on the evaluation measure are provided in Appendix A. The baseline is the original nnU-Net, with and without the use of random on-the-fly data augmentation (both iconographic and spatial, as described in Section 3.1.2).

First of all, the results show that the use of iconographic data augmentation helps the networks to reduce the false negative voxels but it does not increase significantly the performance. This may be due either to the fact that the contrast variability in these structures is already representative in the training set or that such augmentations are not sufficient to significantly improve the results. The problem related to contrast heterogeneity is addressed in the Chapter 4. By contrast, the use of spatial data augmentation or pose and size homogenization is critical for improving performance, particularly on tumor segmentation. Our proposed STN^1 to homogenize pose and size outperforms (better mean and decrease of the standard deviation for both Dice score and 95HD) both the transfer learning (3D and 2D) results (Table 3.3 and Table 3.4) and the baseline with data augmentation. However, a Wilcoxon Signed-Rank Test between the proposed method and the baseline with data augmentation showed non-statistically significant improvements for Dice score for no value of alpha, intuitively noticeable by the still high standard deviations, while the improvement was significant for the 95HD of kidney segmentation with $\alpha = 0.10$, confirming an interesting reduction in local errors. In addition, the total training time (both STN and nnU-Net) is less (8h difference) than the

Table 3.7: 2D Results (mean and standard deviation of Dice score, precision, recall, 95th percentile of Hausdorff distance (95HD) and total training time) on the Necker PRAC database, adding the proposed STNs to the baseline nnU-Net (without data augmentation). DAI = iconographic data augmentation; DAS = spatial data augmentation; TRN = training; S = structure; K = kidneys; T = renal tumors.

Architecture	DAI	DAS	TRN Time	S	Dice Score [%]	Precision [%]	Recall [%]	95HD [mm]
nnU-Net			22h	K	86.77 (8.39)	89.74 (4.62)	84.73 (12.03)	12.84 (20.98)
				T	74.07 (23.34)	86.01 (22.67)	67.43 (24.86)	19.20 (20.97)
nnU-Net	✓		23h	K	88.61 (6.26)	89.94 (4.65)	87.52 (8.41)	12.18 (20.51)
				T	75.84 (24.72)	84.22 (25.96)	71.01 (25.92)	18.53 (23.88)
nnU-Net	✓	✓	33h	K	89.37 (4.76)	89.51 (5.23)	89.57 (6.69)	11.12 (15.86)
				T	82.93 (18.55)	90.11 (15.12)	79.82 (19.73)	16.06 (22.79)
STN pose-size + nnU-Net			25h	K	88.91 (6.35)	89.45 (3.05)	88.79 (9.65)	11.85 (19.24)
				T	81.43 (24.04)	83.40 (26.03)	80.92 (23.73)	22.78 (32.12)
STN pose-size + nnU-Net	✓		25h	K	89.53 (4.44)	90.99 (2.61)	88.40 (7.37)	8.64 (16.55)
				T	83.19 (17.31)	84.41 (19.83)	82.47 (20.72)	15.51 (21.67)

training time from the nnU-Net with complete data augmentation. Such tests confirm that with proper spatial homogenization of data, similar or better results can be obtained compared to the use of costly (in terms of memory and time) data augmentation, taking advantage of the network’s ability to learn general and relative positions of structures to be segmented, which are meaningful in the medical field. Furthermore, the spatial data augmentation that experimentally showed the greater advantages to the baseline nnU-Net using it as only data augmentation technique, was the use of mirroring. This cannot be reproduced by our module by choice confirming how spatial relationships and absolute positions become probably more useful for the network to achieve satisfactory results.

Finally, for this combination, it was noted that end-to-end training brings no benefit, slightly reducing the performance of both networks.

Table 3.8: Results (mean and standard deviation of Dice score and total training time) on the Necker PRAC database reducing the size of the input image for nnU-Net (memory allocated column refers only to nnU-Net, STNs occupy less than 4GB of RAM in the GPU also with 512×512 inputs). Note that each network is trained individually. DAI = iconographic data augmentation; DAIS = spatial + icon. data augmentation; S = structure; K = kidneys; T = renal tumors.

Architecture	Input size U-Net	Training Time	Memory allocated	S	Dice Score [%]	Precision [%]	Recall [%]	95HD [mm]
nnU-Net (+ DAIS)	512×512	33h	10.05GB	K	89.37 (4.76)	89.51 (5.23)	89.57 (6.69)	11.12 (15.86)
				T	82.93 (18.55)	90.11 (15.12)	79.82 (19.73)	16.06 (22.79)
STN pose-size + nnU-Net (+DAI)	512×512	33h	10.05GB	K	89.53 (4.44)	90.99 (2.61)	88.40 (7.37)	8.64 (16.55)
				T	83.19 (17.31)	84.41 (19.83)	82.47 (20.72)	15.51 (21.67)
STNp-s + STNcrop + nnU-Net (+DAI)	512×512	28h	10.05GB	K	89.22 (5.47)	91.30 (2.38)	87.58 (8.65)	4.66 (2.54)
				T	84.31 (15.91)	86.90 (16.32)	82.81 (16.20)	9.90 (9.40)
STNp-s + STNcrop + nnU-Net (+DAI)	256×256	19h30	3.52GB	K	88.89 (9.19)	90.81 (3.94)	87.86 (13.32)	8.35 (12.45)
				T	81.24 (17.80)	88.45 (14.58)	77.55 (21.41)	17.80 (23.46)

The combination of the two *STNs* with images at original size leads to more improvements in performance to tumor segmentation compared to using *STN*¹ alone and nnU-Net with data augmentation. Moreover, the resampling of the bounding box of *STN*² at 256×256 leads to a gain in time and requested memory as shown in Table 3.8, while maintaining high performance.

This is due to the fact that the U-Net has a smaller image as input and one layer less in depth. In this case the drop in performance depends on the renal tumor size, and consequently on the size of the ROI, which varies from 128×128 to 380×380 . This means that, when reducing the input size of the U-Net to 256×256 , we actually downsample the ROI thus losing important information, as shown in the last row of Figure 3.11. As for the previous combination, the Wilcoxon Signed-Rank Test showed that there are non-statistically significant improvements except for the 95HD of kidney segmentation, confirming the ability of the proposed method to reduce local errors, in particular when the input size is kept to 512×512 , and to maintain similar performance to the baseline, when the input size is reduced to 256×256 . We believe that the proposed differentiable module to localize and resample ROIs may be important for other datasets with smaller and more regular structures to segment compared to the size of the image, or when training time and memory are limited. It is important to emphasize that to overcome the errors found for STN^2 and detailed in the previous paragraph, we used a “security-margin” system adding 10% of input image in scaling parameters calculation of θ^2 . Some later experiments have shown that this step can be overcome with end-to-end training, which avoids accumulation errors on the borders.

In Figure 3.11, results of the proposed network are illustrated step-by-step on 512×512 images. In the first four rows, we do not change the input size of the U-Net, whereas in the last row we reduce it to 256×256 . This results in a less detailed image and thus a drop in performance.

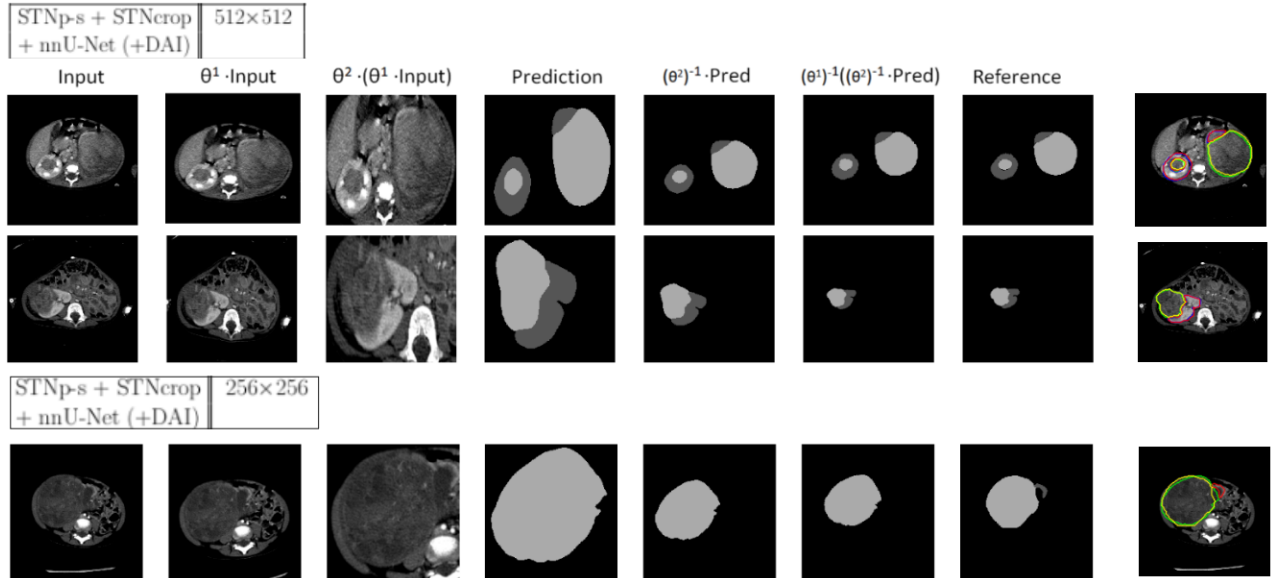


Figure 3.11: Qualitative results of our method illustrated step-by-step on children images. In the last line, the cropped image is downsampled to 256×256 and it can be noticed that the boundaries between tumor and renal cavities are lost. Kidney in dark gray and renal tumor in light gray. On the last column on the right, the overlapping segmentation for the reference (kidneys in red and tumor in green) and for the prediction (kidneys in blue and tumor in yellow).

It is interesting to note that by applying this method to adult images, with an already almost homogeneous database in pose and size, the application of STN_1 or both $STNs$ did not lead to any improvement in performance, and the results are comparable (e.g. for both $STNs$ we have: DS kidney = 93.68 (4.30) and DS tumor = 62.35 (33.80), see Table 3.1 for comparison with state-of-the-art methods). It is also interesting to note that a zoom on the bounding box

had also been proposed by some of the networks at the KiTS19 [53] and KiTS [52] challenges, attaching two U-Nets in cascade: the first one gives a rough estimation of the segmentation that is later used to extract smaller patches centered on the first predicted segmentation; the smaller patches are given as input to the second network for the final segmentation. This idea is feasible in adults since, as shown previously, the tumor is small and does not change the renal parenchyma, thus the bounding box is almost constant. Despite this technique, the nnU-Net still ranked first in KiTS19 [53] using large patches as described in Section 3.1.2. This means that the difficulties in adult tumor segmentation could be more related to not well-defined contours or contrast heterogeneity than to their size, as visible in Figure 3.12.

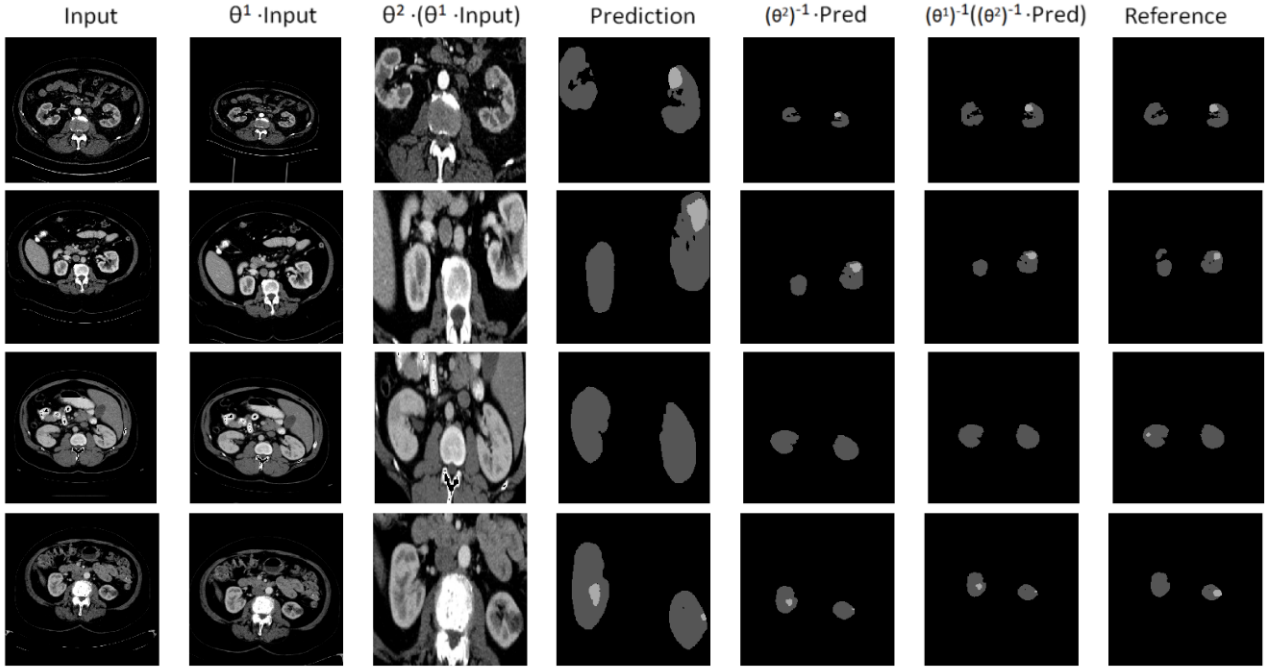


Figure 3.12: Qualitative results of our method illustrated step-by-step on adults images. It can be noticed how renal tumors are very small and have not well-defined contours. Kidney in dark gray and renal tumor in light gray.

3.5 Transfer learning with *common* size and pose

Spatial transformations are necessary to make transfer learning between adults and children possible, as explained in Section 3.3. To achieve this goal, our first idea is to use STN to homogenize adults and children datasets in size and pose using a reference image, as presented in Figure 3.13.

The method proposed in Section 3.4.1 can be used in its combination $STN^1 + U\text{-Net}$. Furthermore, iconographic data-augmentation can be added to obviate possible differences in contrast and brightness between populations. Before being given as input to the U-Net, the children dataset is spatially transformed in pose and size using the STN to be as similar as possible to a reference pose, created from a sample image of the adults dataset. As shown in the previous sections, adult images in the database already exhibit homogeneous pose and

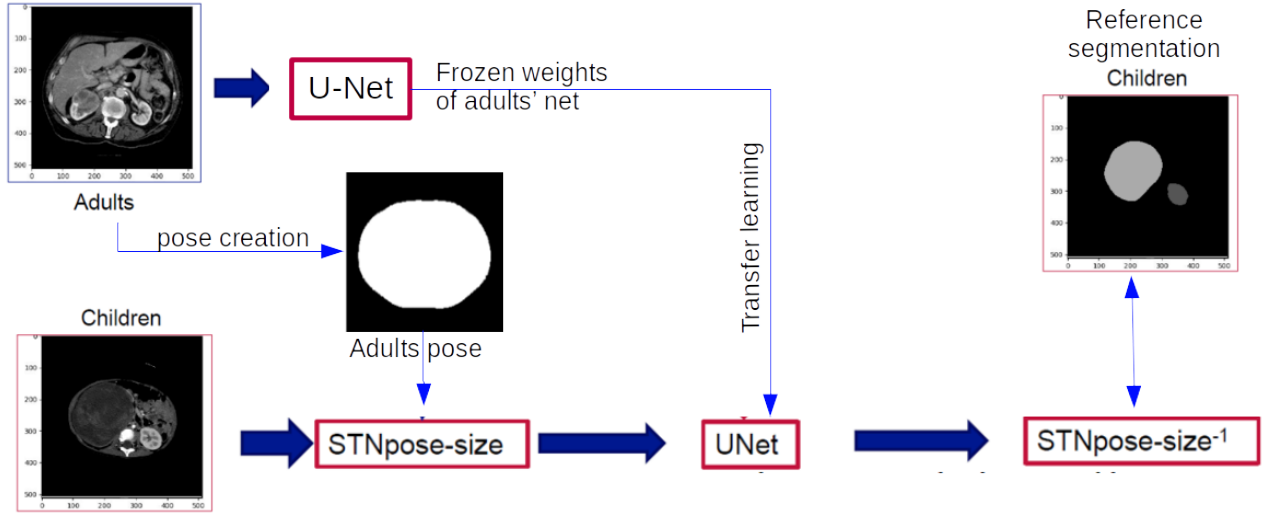


Figure 3.13: Framework for the proposed method of common size and pose for adults and children images. The U-Net can be the same as for adults or its weights can be used as initialization of the network of children.

size. So we can continue to operate using only adults' net weights and not the adults' dataset (as in the more realistic case where images are not available), as discussed in Section 3.3. We will refer to the reference adult pose as *common* reference pose.

We tested the proposed approach both using the nnU-Net trained only on adults in direct inference on pediatric images, and fine-tuning the nnU-Net with a part of the available ceCT images of children. In fact, we put ourselves in the more classic case where a hospital or institution does not have many subjects to train a network with good performance, as shown in Figure 3.5 in the case of only 25 subjects for training (and validation) and 15 as test set.

Results of this technique using direct inference are shown in Table 3.9. The data-sets are the same as the ones presented in Section 3.3.

Table 3.9: Results (mean and standard deviation of Dice score) on 15 children test patients with and without STN for homogenization at a *common* pose and size with adults. Results are obtained with the same pixel size than the one of adults, and with a smaller pixel size of $0.46 \times 0.46 \text{ mm}^2$ using weights of 2D nnU-Net trained on adults KiTS19 database [53] with pixel size of $0.78 \times 0.78 \text{ mm}^2$.

Technique (2D Networks)	Children pixel size (mm^2)	Dice Score Kidney	Dice Score Tumor
Direct Inference (weights frozen)	0.78×0.78	37.43(29.95)	36.88(24.91)
Direct Inference (weights frozen)	0.46×0.46	48.51(26.33)	42.75(33.01)
STN^1 + Direct Inference (weights frozen)	0.78×0.78	50.37(19.03)	28.26(23.22)
STN^1 + Direct Inference (weights frozen)	0.46×0.46	53.43(19.66)	33.00(23.84)

The use of STN^1 for a *common* pose and size leads to improvement of the segmentation of the kidneys, while it shows worse results for that of the tumors. The use of a smaller pixel size for children, which makes images more similar to those of adults, facilitates the task for both the pre-trained U-Net and the STN, thus providing better results. The worst performances in renal tumor segmentation are probably due to the fact that children's tumors become even bigger and therefore even more different than those of adults.

Then we put ourselves in the event that in addition to the 15 test subjects, other 25 subjects are also available. In this scenario, we test a fine-tuning strategy of the entire network. Moreover, starting from what was seen in the previous analysis and shown in Table 3.9, a

smaller pixel size equal to $0.46 \times 0.46 \text{ mm}^2$ is used for children. Results are shown in Table 3.10.

Table 3.10: Results (mean and standard deviation of Dice score) on 15 children test patients of networks fine-tuned with 25 children patients and adults’ weights as initialization. STN^1 for homogenization at a *common* pose and size with adults. All children pixel size is equal to $0.46 \times 0.46 \text{ mm}^2$. Last row shows results using the entire database of 65 children patients. DAI = iconographic data augmentation; DAS = spatial data augmentation; S = structure; K = kidneys; T = renal tumors.

Technique (2D networks)	lr	S	Dice Score [%]	Precision [%]	Recall [%]	95HD [mm]
Direct training (+DAIS) (no fine tuning)	0.01	K	87.41 (6.24)	85.15 (8.01)	90.01 (5.10)	14.19 (26.78)
		T	78.50 (20.55)	91.34 (14.08)	72.54 (20.77)	18.89 (34.82)
Fine-tuning (+DAIS) (weights as init.)	0.001	K	87.66 (6.39)	85.69 (7.56)	89.98 (5.80)	12.15 (24.82)
		T	80.12 (17.63)	91.05 (14.84)	75.15 (17.40)	17.88 (34.44)
Fine-tuning (+DAIS) (weights as init.)	0.01	K	86.99 (7.24)	84.43 (9.26)	89.79 (6.02)	14.41 (28.28)
		T	81.07 (16.56)	90.72 (14.81)	76.44 (16.00)	16.89 (29.39)
STN^1 + Fine-tuning (+DAI) (weights as init.)	0.001	K	87.79 (5.58)	86.12 (8.09)	89.59 (4.43)	11.92 (26.40)
		T	80.39 (22.27)	87.19 (14.26)	78.37 (21.81)	18.53 (34.96)
STN^1 + Fine-tuning (+DAI) (weights as init.)	0.01	K	88.05 (5.90)	85.82 (8.82)	90.36 (4.58)	12.23 (26.99)
		T	82.98 (14.88)	89.29 (13.83)	80.69 (14.04)	16.14 (21.62)
Direct training (+DAIS) using all 65 patients (no f.t.)	0.01	K	89.37 (4.76)	89.51 (5.23)	89.57 (6.69)	11.12 (15.86)
		T	82.93 (18.55)	90.11 (15.12)	79.82 (19.73)	16.06 (22.79)

In this case, the use of fine-tuning alone does not bring much improvement over training using only the 25 subjects, even with a smaller pixel size to make the images more similar to those of adults and a strong spatial augmentation, confirming what was discussed in Section 3.3. Instead, the use of STN^1 to homogenize to the *common* pose allows for images more similar to those used for training the adults’ network, leading to a boost to the results, with a large decrease in standard deviation. It is also interesting to note that a higher learning rate is needed for the best results. Such performances are comparable to 2D training using the totality of 65 patients of the pediatric training database. Thus, we can assess that the use of a homogenization system is more effective than strong spatial augmentation in the case of transfer learning from weights trained on adults to a limited pediatric database.

3.6 Conclusion

In this chapter we first analyzed the no-new-U-Net as high performance network for medical image segmentation and in particular as winner of the Kidney and Tumor Segmentation (KiTS) challenge of 2019 focus on adults images. Given the lack of work on pediatric images, the in-depth analysis of this method seemed to us an interesting starting point. It is important to point out that this chapter represents the study carried out along about the first year of doctoral studies, and for this reason it is focused on the KiTS19 Challenge and Database and not on the KiTS21. Nevertheless, the KiTS challenge of 2021 does not have major differences on the database and the best performing methods are still based on nnU-Net.

The key points of this method were found to be the image pre-processing and network planning, which we subsequently used for the other segmentation experiments that are shown in the rest of this manuscript. As far as training is concerned, each of the techniques that are used in nnU-Net (use of the bias, poly learning rate policy, etc.) resulted in a contribution to

the final performances. Major improvements were observed thanks to the use of oversampling due to the strong imbalance of the target structures in 2D slices or 3D patches (recurring problem in medical images) and the use of the spatial and iconographic data augmentation due to medical databases that are often limited in number of patients.

This last point was of interest to us, in fact the spatial heterogeneity of medical databases, in particular pediatric ones, is very high and this leads to using a large data augmentation that requires a lot of memory and time. We proposed instead the use of a new homogenization techniques using Spatial Transformer Network in order to reduce data variability in size and pose, in place of enlarge this via data augmentation. Our idea is also related to the ability of U-Net networks to learn the absolute and relative position of the structures under examination. Through the analysis of the results, we can conclude by stating that our proposition is effective, improving performances and computational time with respect to standard data augmentation. We also believe that these results stand also as a criticism of the disproportionate use of this technique that is sometimes seen in the literature.

We also proposed the use of a second STN to crop images around the structures to segment. This can improve the results thanks to a zooming on target structures or can save even more computational time and memory, while maintaining high performance, thanks to an actual cropping. This can be really useful in case of limited training time and memory available. However, this method presents some problems, related for example to the fact of having to choose a size in which to sample the cropped image that has to be the same for all images.

The results shown for the segmentation are focused on 2D networks. This is due to the fact that 3D STN networks require the entire volume as input (and not just the 2D slices), drastically reducing the number of samples available, thus limiting performance and their consequent use in segmentation networks. Besides, the memory required to use an entire 3D image can be an additional problem. Moreover, the homogenization of size and pose is however limited to the whole body and not to individual organs. In fact, if on 2D this homogenization is sufficient to improve the performance of the network, in the application to segment 3D volumes this would not entail major advantages due to the lack of spatial consistency between the organs of the abdominal area. This point is clarified in the next chapter.

Furthermore, despite the high results achieved in 2D, the method we propose is less performing than the use of 3D patches on a classic nnU-Net, thus a 3D network with strong spatial and iconographic data augmentation. In fact, as described at the beginning of this chapter, 3D segmentation methods are more efficient thanks to the greater anatomical information of the 3D patches compared to 2D slices. For this reason, in the case of segmentation task, we focused on 3D networks in the rest of doctoral studies.

Eventually, in this chapter we have also examined the transfer learning between adults and children both affected by renal tumors which, however, are of different sizes between the two categories of subjects. We have shown that the classical transfer learning methods, such as fine-tuning, from adults' pre-trained weights do not lead to further improvement of the performances. To make transfer learning effective, spatial transformations are necessary. We showed that the use of STN to homogenize size and pose is effective in increasing the performance of transfer learning, making the images of the target network more similar to those of the source network.

Chapter 4

Cross-domain CT image translation using CycleGAN

Heterogeneity in contrast is one of the major difficulties in medical image segmentation when using Convolutional Neural Networks (CNN), in particular in contrast-enhanced Computed Tomography (ceCT) images. As explained in the Introduction 1, the effect of the contrast agent on the pixel intensity is not always the same among patients due to different factors, such as acquisition times and patient morphology. Furthermore, the presence of a tumor or thrombosis in the vessels can also cause heterogeneity in contrast within an anatomical structure. This raises difficulties during segmentation, and manual corrections are often needed.

In [119, 125, 158] the authors show that the combined use of ceCT and contrast-free (CT) CT images is able to deal with the heterogeneity of ceCT images and thus improves segmentation. However, in order to limit ionising radiations, clinicians often acquire only one CT modality. One common computational approach to compensate for the absence of an imaging modality is to use generative models [23, 149] to synthesise it. In the absence of paired data sets, unsupervised translation methods, based on CycleGAN [157] and UNIT [87], have been proposed [29, 71, 107, 149]. Recently, some authors have already considered applying CycleGAN [119, 125] or UNIT [158] to artificially remove or add contrast medium on CT images.

CycleGAN [157] is an evolution of Generative Adversarial Network (GAN) [47], which introduces a second neural network that tries to solve the inverse task, namely reconstructing the input, as illustrated in Figure 4.1. A cycle consistency loss function (detailed later on, in Equation 4.7) is combined to the adversarial loss function to overcome the lack of paired data.

UNIT [87] is another model conceived for the unpaired setting. This generative model is composed of two variational autoencoder networks, which work on two different domains but share the same latent space. Different modifications have already been proposed for both methods, such as the use of Wasserstein distance [2], attention mechanisms [39, 72] and U-Net as discriminator network [120]. In [?], the two models are compared to a simple GAN, to transform unpaired MR brain images into CT images and vice versa. The simple generator is not able to produce images that are as realistic as the ones generated using CycleGAN or UNIT. However, these models do not guarantee to preserve fine structures [158] and may produce artefacts [121, 149], which prevent their use for the segmentation of small and heterogeneous structures, such as blood vessels. In particular, the cycle consistency loss function enforces a relationship only at a distribution level. In [103], the authors demonstrate that CycleGAN can

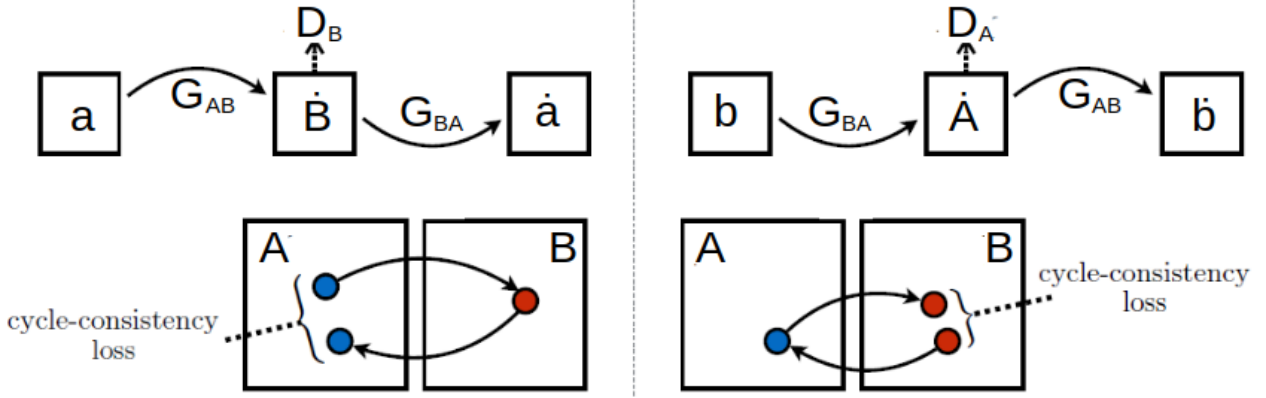


Figure 4.1: Diagram of how cycle loss works. Left: forward Cycle-Consistency. Right: backward Cycle-Consistency. A and B are the two domains describing two set of images, a and b are single images extracted from the respective domains, G_{AB} and G_{BA} are the generators, D_A and D_B the discriminators, $\hat{b} = G_{AB}(a)$ and $\hat{a} = G_{BA}(b)$, $\hat{a} = G_{BA}(\hat{b})$ and $\hat{b} = G_{AB}(\hat{a})$. Adapted from [157].

deliver ambiguous solutions, especially for substantially different distributions as in medical imaging. Several works tried to address this limitation by adding more terms to the loss function, such as mutual information [42] and a perceptual loss term [3, 146], that require no supervision. Despite the different methods proposed, the anatomical constraint remains insufficient.

As a matter of fact, another challenge when dealing with unpaired 3D medical images is the lack of 3D consistency. With current hardware memory limitations, it is difficult to train a 3D network taking as input a whole 3D volume. Instead, a common approach is to use 2D networks that take a slice of a 3D volume along one axis. Moreover, in the unpaired scenario, we can have different numbers of slices for the same anatomical region among patients, leading to difficulties to select anatomically-paired slices. In fact, in [121], authors showed that it is fundamental to inform both the generator and the discriminator on the specific regions that should be affected by the contrast materials. For these reasons, they proved that the use of paired data (albeit slightly misaligned because ceCT is acquired some seconds after CT) is more effective than unpaired data [121]. Some authors [147] claim that the use of unpaired data can be mitigated by exploiting the approximately common anatomy between subjects. They refer to this as *position-based selection (PBS)* strategy. However, this method has been proved to perform on par with the use of paired subject only on brain images and for MR-CT translation [147].

In fact, first, in the abdominal region, the different sizes and lengths of the organs must be taken into account, implying that the slice a of the patient i with \mathcal{D}_a slices may not have the same anatomical content as slice $b = a \cdot \frac{\mathcal{D}_b}{\mathcal{D}_a}$ of patient j with \mathcal{D}_b slices. Eventually, the use of 3D affine registration could be a solution to the problem, but the difference between the two domains, the difficulty of identifying the fixed reference image, and the high variability in shape and relative size and pose of abdominal organs among subjects (especially in 3D) may lead to misalignment.

Then, differences between ceCT and CT domains are more subtle than for example between MR and CT or PET and CT. Due to the physical differences in acquisition, these modalities exhibit important differences in texture which ease critic mechanisms of discrimination. Conversely, ceCT and CT images are distinguishable only in certain anatomical parts and only in

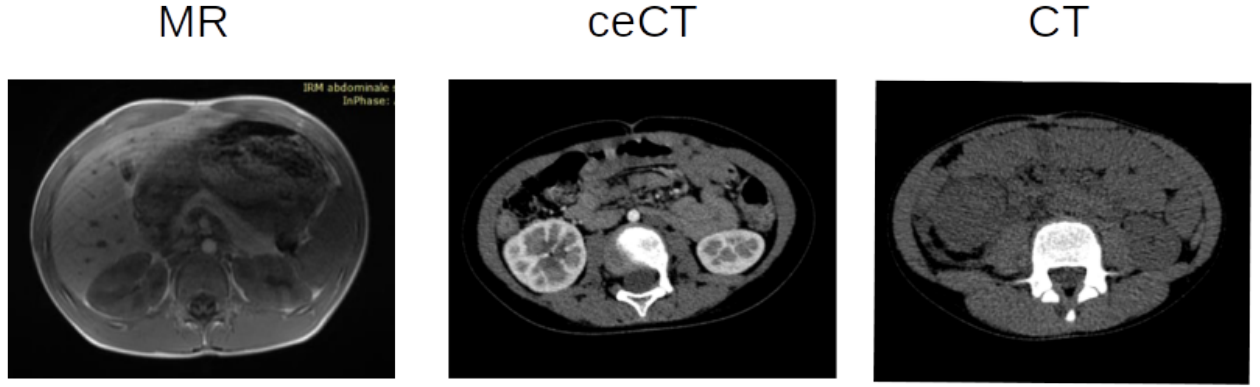


Figure 4.2: Three slices examples of MR, ceCT and CT images from the Necker database. The MR image presents different texture from ceCT and CT images, which instead differ only in certain areas of the image.

some 2D slices. Figure 4.2 shows what is claimed.

To address these issues, we propose an extension of the CycleGAN method which includes:

1. the automatic selection of the region of interest by exploiting anatomical information, in order to reduce the anatomical distribution of 3D data acquired with different fields of view;
2. the use of a *Self-Supervised Body Regressor* (SSBR), adapted from [144], to select anatomically-paired slices among the unpaired ceCT and CT domains, and help the discriminator to specialize in the task;
3. the use of the SSBR score as an extra loss function that constrains the generator to produce a slice describing the same anatomical content as the input, inspired from the auxiliary classifier GAN [104];
4. the use of the input image as a template for the generator, as in [19], and the use of an anatomical binary mask to constrain the output.

The proposed method is generic and could be used in different medical applications, i.e. different body regions such as brain or lungs, or different translation modalities such as MRI to CT or T1-w to T2-w. Here, we propose to use it for the generation of CT abdominal images from ceCT images and vice versa. In addition, the proposed method is applicable regardless of the generating network G and discrimination mechanism D chosen. For this reason, an initial state-of-the-art assessment is conducted in order to find the best combination of G and D for our specific task. We test also the use of a generated modality, in combination with the complementary original one, to improve segmentation performance on blood vessels of pathological patients. To the best of our knowledge, this strategy has never been tested for such an application.

We show that our method greatly improves the ceCT-CT translation quality compared to state-of-the-art methods. As a consequence, the segmentation performances using generated images are also improved, achieving both qualitative and quantitative results comparable to the ones using both real images. The proposed translation method led to a paper accepted at the Colloque Francophone de Traitement du Signal et des Images (GRETSI) 2022 [GRETSI-22]. Its extended version (with improvement on SSBR training, more exhaustive qualitative

and quantitative evaluation, application on blood vessel segmentation) has recently been accepted at the British Machine Vision Conference (BMVC) 2022 [BMVC-22].

It is important to highlight that, in our proposition, the use of synthetic images is intended to increase segmentation performances and not to be used for clinical diagnosis, as in [76].

It should also be noted that in our work we focus on CNN-based methods, as state-of-the-art methods for unsupervised medical image translation. While interesting works on image-to-image translation based on Vision Transformer (ViT) are starting to be explored [68, 73], their application in the medical domain is limited due to the restricted number of data available. For this reason, existing works focus only on paired medical data sets [27]. ViT architectures address one of the problem of CNN-based methods: performances in learning long-range dependencies are limited to their localized receptive fields [49]. ViT encodes images as a sequence of 1D patch embeddings and utilizes multi-head self-attention modules to learn a weighted sum of values that are calculated from hidden layers. This results in differentially weighting the significance of each part of the input data and effectively learning the long-range information [49]. However, not everyone agrees on Transformer’s revolution. Some authors claim that CNN, with some single self-attention modules (e.g. MLP) [86, 131] or a big receptive field that can capture global context [113, 116], exhibit both locality and spatial invariance after training. Moreover, other authors [88, 141] proved that an up-to-date CNN, such as a Res-Net, with the same number of parameters than a ViT can achieve same or even better performances. Finally, I would like to end with a personal thought, in accordance with what stated in [88]. Investigating those model designs inevitably results in an increase in carbon emissions (more powerful GPUs, more experiments to understand, more training hours) and we should not do it if we can achieve already satisfactory performance with less carbon-demanding methods. Nevertheless, for the sake of completeness, we test a Transformer-GAN, namely TransGAN [68], on our unsupervised medical task, using the same technical specifications as the other networks (details in Section 3.4.1), and thus reducing the number of parameters of the ViT to fit our conditions (further details in Section 4.2.1).

The chapter is structured as follows. In Section 4.1 we describe our proposed methods for the input selection via SSBR and for the anatomically constrained CycleGAN. In Section 4.2, first we details the implementation of our algorithms, then qualitative and quantitative results are discussed. Eventually, we show blood vessel segmentation results when using a generated CT modality instead of the real one. Section 4.3 presents our conclusion on this contribution and the next steps planned.

4.1 ceCT-CT image translation using CycleGAN with anatomical constraints

As stated in the previous section, when using 2D unpaired medical data, the selection of consistent (*i.e.*, corresponding to the same region of interest (ROI)) and anatomically similar slices between the two domains is very important to facilitate the generative process. To this end, we propose to leverage a Self-Supervised Body Regressor (SSBR) [144], a CNN that finds common features on anatomically similar slices from unlabeled CT images. This results in assigning the same label for slices describing the same anatomy while belonging to different patients. The SSBR is trained to estimate slice scores which are monotonous functions of the slice indices. However, there is no guarantee to obtain the same range of scores for different

modalities. We propose a solution to this problem. The method described in this section is summarized in Figure 4.3.

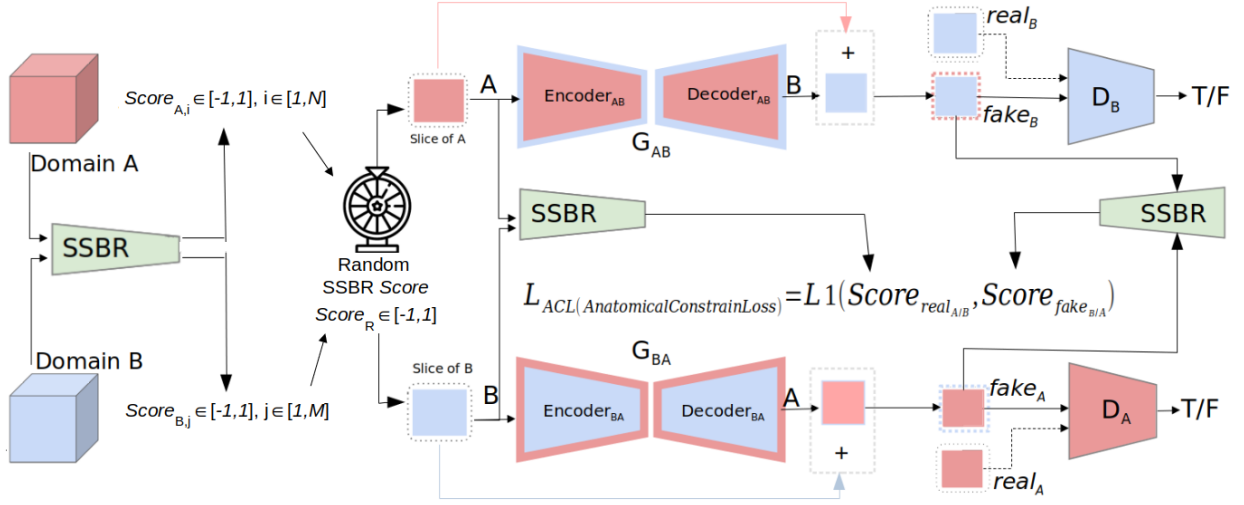


Figure 4.3: Proposed method for the selection of anatomically-paired slices via *Self-Supervised Body Regressor*, and its use as a loss function L_{ACL} . $Score_{A,i}$ is the value assigned in $[-1, 1]$ by our proposed pre-trained *SSBR* for each slice i in $[1, \mathcal{D}_A]$ for an image of the domain A , while $Score_{B,j}$ for each slice j in $[1, \mathcal{D}_B]$ for domain B . $Score_R$ is a random score extracted in $[-1, 1]$ used in the anatomically-paired selection process detailed in the text. The arrows ending with the “+” sign indicate that the input is added to the output of the network, producing $fake_{A/B}$. The $Score$ of the latter is inferred by the *SSBR* and compared with the $Score$ of the $real_{B/A}$ input via L_{ACL} loss function.

4.1.1 Input selection via SSBR

First of all, because of the different fields of view (FOV) in the two datasets, it is important to select an appropriate ROI. This can be done off-line and manually, as in [119, 158]. Here, instead, we first propose a simple, automatic and on-line method to select only slices from the abdominal region. We automatically select the first slice of the lungs and the last slice of the intestinal area as upper and lower landmarks, which is easy due to the strong presence of black pixels in both ceCT and CT acquisitions.

Then, to select anatomically-paired slices we propose the use of an *SSBR*, as shown in Figure 4.3, instead than PBS strategy as in [147], where a slice a is selected from a patient of domain A with \mathcal{D}_A slices and a slice $b = a \cdot \frac{\mathcal{D}_B}{\mathcal{D}_A}$ is selected from a patient of domain B with \mathcal{D}_B slices. For training, we consider the domain Ω (of which A and B are sub-domains), from which at each iteration K patients are selected. Once the automatic restriction to the renal ROI is applied, J slices are extracted per patient. The first and last slices of the ROI are always selected while the others are randomly selected between these two. We optimize three loss functions that do not require annotated anatomical labels. The first one, as in [144], favors an increasing order of *SSBR* scores according to the positions of the slices, avoiding repeating scores and ensuring similar scores for adjacent regions:

$$L_{order} = - \sum_{k=1}^K \sum_{j=1}^{J-1} \log(s(Score_{k,j+1} - Score_{k,j})) \quad (4.1)$$

where $Score_{k,j} \in [-1, 1]$ is the SSBR output for slice j of CT volume k , s is the sigmoid activation function, K is the number of CT volumes in the chosen set (mini-batch) and J is the number of slices in each volume, as mentioned above.

The second loss function exploits the automatic selection of the ROI, forcing the first and last slices to have a score of -1 and $+1$ respectively:

$$L_{norm} = \sum_{k=1}^K (c(Score_{k,1} + 1) + c(Score_{k,J} - 1)) \quad (4.2)$$

where c is a smoothed L1 norm. This function guarantees the same score range for both modalities.

The third loss function takes into account the anatomical variability of the abdominal area. Using the binary mask BM of the body for each slice (easily obtained in CT), we want the difference between successive scores to be an increasing function of the normalized cardinality of the intersection of the BM of successive slices ($BM_{k,j-1}$ and $BM_{k,j}$ for volume k):

$$L_{anat} = \sum_{k=1}^K \sum_{j=1}^{J-1} c(\Delta_{k,j+1}^{BM} - \Delta_{k,j}) \quad (4.3)$$

with $\Delta_{k,j}^{BM} = 1 - \frac{|BM_{k,j} \cap BM_{k,j-1}|}{|BM_{k,j-1}|}$ and $\Delta_{k,j} = Score_{k,j} - Score_{k,j-1}$

This is done in order to increase the difference in score between slices with higher anatomically difference and not fall into the trivial linear solution.

Eventually, the terms of the cost function are combined by a weighted average, and the function to be optimized is:

$$L_{SSBR} = w_o L_{order} + w_n L_{norm} + w_a L_{anat} \quad (4.4)$$

where w_o , w_n and w_a are empirically chosen weights that balance the three loss terms.

Once the SSBR is properly trained by optimizing Equation 4.4 (details in the next section), to extract the anatomically-paired slices for each iteration of the CycleGAN we do the following:

1. A single patient is selected for each of the unpaired ceCT and CT domains, called domains A and B ;
2. The 3D volumes are automatically restricted to the abdominal region;
3. SSBR scores are predicted for each 2D slice of the two 3D ROIs, using the pre-trained SSBR;
4. N random SSBR scores, denoted by $Score_{\mathbf{R}_n}$, are sampled in $[-1, 1]$, where N is the selected number of slices corresponding to the size of the mini-batch;
5. For each $Score_{\mathbf{R}_n}$, the slice with the closest score is selected in each domain, as $\arg \min_d |Score_{\mathbf{R}_n} - Score_{\cdot,d}|$ where \cdot is the domain (A or B) and d is the selected slice in $[1, \mathcal{D}_A]$ for A and $[1, \mathcal{D}_B]$ for B .

4.1.2 Anatomically constrained CycleGAN

Inspired by [104], we propose the use of the pre-trained SSBR as an auxiliary classifier to enforce the anatomical consistency (i.e., same body parts) between the input and the synthesized output. During the training phase of the generator, we add to the loss functions of the standard CycleGAN an L1 norm between the SSBR score of the input *real A* (resp. *real B*) and the SSBR score of the generated slice *fake B* (resp. *fake A*), called the *Anatomical Constraint Loss (ACL)*:

$$\begin{aligned} L_{ACL}(G_{AB}, G_{BA}) &= \frac{1}{N} \sum_{n=1}^N |Score_{real_{A/B},n} - Score_{fake_{B/A},n}| = \\ &= \mathbb{E}_{a \sim p_{data}(a)} [||Score_{G_{AB}(a)} - Score_a||_1] + \mathbb{E}_{b \sim p_{data}(b)} [||Score_{G_{BA}(b)} - Score_b||_1] \end{aligned} \quad (4.5)$$

where the two generators are denoted by G_{AB} and G_{BA} , the two sets of images are described by domain A and domain B (sub-domains of Ω , as domain of abdominal CT images), a and b are single images, N is the mini-batch size, and the probability distribution for each domain is denoted by $p_{data}(a)$ and $p_{data}(b)$, respectively. The L_{ACL} loss function constraints the generator to produce highly detailed slices describing the same anatomical region as the input, in order for the SSBR to produce the same score as this.

Therefore, the original loss function of CycleGAN [157] is modified as:

$$\begin{aligned} L(G_{AB}, D_B, G_{BA}, D_A) &= L_{GAN}(G_{AB}, D_B, G_{BA}, D_A) + w_{cyc} L_{cyc}(G_{AB}, G_{BA}) \\ &\quad + w_{idt} L_{idt}(G_{AB}, G_{BA}) + w_{ACL} L_{ACL}(G_{AB}, G_{BA}) \end{aligned} \quad (4.6)$$

with

$$L_{cyc}(G_{AB}, G_{BA}) = \mathbb{E}_{a \sim p_{data}(a)} [||G_{BA}(G_{AB}(a)) - a||_1] + \mathbb{E}_{b \sim p_{data}(b)} [||G_{AB}(G_{BA}(b)) - b||_1] \quad (4.7)$$

and with

$$L_{idt}(G_{AB}, G_{BA}) = \mathbb{E}_{a \sim p_{data}(a)} [||G_{BA}(a) - a||_1] + \mathbb{E}_{b \sim p_{data}(b)} [||G_{AB}(b) - b||_1] \quad (4.8)$$

The generator G_{AB} transforms domain A in B , G_{BA} does the inverse process, D_A is the discriminator for domain A and D_B for domain B (involved in L_{GAN} , detailed below), as shown in Figure 4.3. The parameter w_{cyc} is the weight for the cycle loss, w_{idt} is the weight for the identity loss and w_{ACL} for the anatomical constrain loss. The advantages of using L_{cyc} and L_{idt} in medical images were demonstrated in [70] by Kang *et al.*. They used a CycleGAN for denoising of coronary CT, and via an ablation study they demonstrated that the use of L_{cyc} avoids creation of any artificial feature that is not present in the input images, while the addition of L_{idt} helps to further discriminate features of interest to avoid hallucinations and also to preserve detailed edge information. We can see the L_{idt} as a regularizer that forces the generator to be near an identity mapping when real samples of the target domain are provided. In this way the model better preserves the content shared between the two domains (e.g. background colors) and will be more conservative for unknown content [70]. Nevertheless, as described at the beginning of this chapter, for more difficult task such as modalities translation task, the L_{cyc} function is not sufficient to achieve desired performances and the regularization done by L_{idt} does not affect the translation performance.

For $L_{GAN}(G_{AB}, D_B, G_{BA}, D_A)$, it is important to clarify that it is composed by two parts, $L_{GAN}(G_{AB}, G_{BA})$ used in the generators optimization and $L_{GAN}(D_B, D_A)$ used for training the discriminators. The optimization of generators and discriminators is done alternately, i.e. for each iteration two cycles are done in which for each one the weights of the part that will not be trained are freezed. We tested two L_{GAN} approaches: the PatchGAN method of the original CycleGAN [157] ($L_{PatchGAN}$) and the Wasserstein loss [2] ($L_{Wasserstein}$), which, according to the authors, favors the stability of the network in the training phase and its convergence.

For the first one, $L_{PatchGAN}$, the loss function is calculated between the output of the discriminator (reduced through convolution to a smaller dimension, named “patch” in [157]) and a matrix composed of ones (\mathcal{M}_1):

$$L_{GAN}(G_{AB}, G_{BA}) = L_{PatchGAN}(G_{AB}, G_{BA}) = \mathbb{E}_{b \sim p_{data}(b)} [\|D_A(G_{BA}(b)) - \mathcal{M}_1\|_2] + \mathbb{E}_{a \sim p_{data}(a)} [\|D_B(G_{AB}(a)) - \mathcal{M}_1\|_2] \quad (4.9)$$

Conversely, the discriminator is trained to assign 1 to real pixels, via \mathcal{M}_1 , and 0 to synthetic ones, through a matrix composed of zero (\mathcal{M}_0):

$$L_{GAN}(D_A, D_B) = L_{PatchGAN}(D_A, D_B) = \mathbb{E}_{B \sim p_{data}(B)} [\|D_A(G_{BA}(b)) - \mathcal{M}_0\|_2] + \mathbb{E}_{a \sim p_{data}(a)} [\|D_A(a) - \mathcal{M}_1\|_2] + \mathbb{E}_{a \sim p_{data}(a)} [\|D_B(G_{AB}(a)) - \mathcal{M}_0\|_2] + \mathbb{E}_{b \sim p_{data}(b)} [\|D_B(b) - \mathcal{M}_1\|_2] \quad (4.10)$$

In this way the generator modifies the gradients of the network to reduce this difference, which indicates that the discriminator is being fooled and interprets the image created by the generator as real.

The Wasserstein loss function $L_{Wasserstein}$ is based on the Wasserstein distance, defined as the shortest average distance necessary to transport one distribution to another one [2]. The two generators are trained to minimize the distance between the distributions of real data and generated images, while the discriminators are trained to maximize it. To this end, the discriminator does not assign a probability, but it scores the realness or fakeness of input data. In order to apply the Wasserstein distance to GAN training, the authors of [2] redefined the objective functions in the following way:

$$L_{GAN}(D_A, D_B) = L_{Wasserstein}(D_A, D_B) = \mathbb{E}_{b \sim p_{data}(b)} [l(D_A(G_{BA}(b)))] - \mathbb{E}_{a \sim p_{data}(a)} [l(D_A(a))] + \mathbb{E}_{a \sim p_{data}(a)} [l(D_B(G_{AB}(a)))] - \mathbb{E}_{b \sim p_{data}(b)} [l(D_B(b))] \quad (4.11)$$

$$L_{GAN}(G_{AB}, G_{BA}) = L_{Wasserstein}(G_{AB}, G_{BA}) = \mathbb{E}_{b \sim p_{data}(b)} [-l(D_A(G_{BA}(b)))] + \mathbb{E}_{a \sim p_{data}(a)} [-l(D_B(G_{AB}(a)))] \quad (4.12)$$

where l is a 1-Lipschitz function. Not being forced to have values between 0 and 1, discriminator weights are clamped within a specific range after each update to prevent very high values. This approximation of Wasserstein distance has been proved to be effective [2, 71, 150, 149]. It is continuous, differentiable and gives a linear gradient, even when the discriminator is well trained. Indeed, in [2] it has been noted that with the original PatchGAN method, the discriminator learned very quickly to distinguish real from fake images, but it then failed to provide useful gradient information to update the corresponding generator. However, the Wasserstein distance is difficult to control and the clamping method can lead to the vanishing gradient problem. A solution is proposed in [48], where instead of applying clamping, a

gradient penalty is implemented in order to punish the discriminator network if the gradient norm deviates from the desired norm of 1:

$$L_{Wasserstein_{GP}}(D_A, D_B) = L_{Wasserstein}(D_A, D_B) + w_{gp}[(\|\nabla_{\hat{\mathbf{a}}} D_A(\hat{\mathbf{a}})\| - 1)^2 + (\|\nabla_{\hat{\mathbf{b}}} D_B(\hat{\mathbf{b}})\| - 1)^2] \quad (4.13)$$

where the last element weighted w_{gp} is the gradient penalty, where $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ are randomly weighted average between a fake and a real sample from each domain as follow:

$$\begin{aligned} \hat{\mathbf{a}} &= v \cdot a + (1 - v) \cdot G_{BA}(b) \\ \hat{\mathbf{b}} &= v \cdot b + (1 - v) \cdot G_{AB}(a) \end{aligned} \quad (4.14)$$

with $v \in [0, 1]$. $L_{Wasserstein_{GP}}$ was used in our tests.

Adding more anatomical constraints Two changes have been proposed to further constrain the model from an anatomical point of view. The first one, as in [19], is the use of the input image as a template on which the generator can work, adding it directly to the output (we will refer to this technique as $I_n A_d$ in images and tables). In this way the two generators do not have to build the image completely from zero, but they focus more on how much and where to change the original image. The output (in $[-1, 1]$) is multiplied by the dynamic of the input database (also in $[-1, 1]$), and in this way a single intensity value can be brought from the minimum to the maximum or vice versa. The second proposed method is done to anatomical constrain the generator even more, using a binary mask BM by which the output is multiplied during inference. In this way only the anatomical structures present in the image are modified. The binary mask is realized thanks to thresholding at the minimum value of the image (corresponding to the background) plus 10% (considering noise). Examples are illustrated in Figure 4.4.

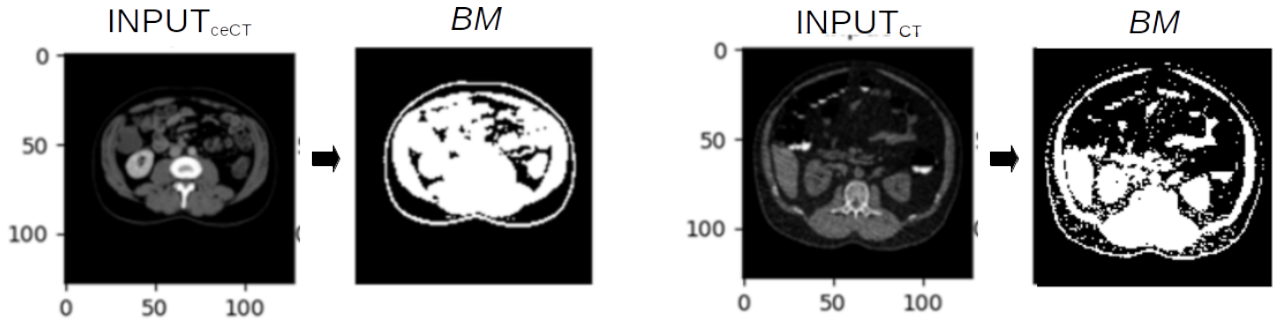


Figure 4.4: Examples of binary mask BM extracted from the input images: Left: ceCT image. Right: CT image.

4.2 Results and discussion

4.2.1 Implementation details

All trainings and tests were run under the technical specifications exposed in Section 3.4.1.

SSBR training For the SSBR, we operated as in [144], using Res-Net-34 [50] as the backbone with an hyperbolic tangent activation function in the last layer. We considered the domain of abdominal CT images Ω , i.e. which contains both ceCT and CT. We set the number of CT volumes K to 32, and the number of J slices extracted from each k to 8. The network is trained for 1000 epochs with Adam [74] optimizer, an initial learning rate (lr) at 0.05 for the first 500 epochs, decreased to 0.01 for the subsequent 250 and to 0.005 for the last 250. The best weights in Equation 4.4 were found empirically as: $w_o = 5 \cdot 10^{-3}$, $w_n = 10$, $w_a = 10$.

Unpaired image translation training The hyperparameters for CycleGAN and UNIT were found empirically on the training set, starting from those in [157] and [87], respectively. The training phase consists of 200 epochs with Adam optimizer and a lr that starts at $2 \cdot 10^{-4}$ and then linearly reduced to 0, from the 100th epoch. The best combination of weights for CycleGAN loss terms was found as 0.5 for w_{idt} , 10 for w_{cyc} and 1 for w_{ACL} . For w_{gp} of $L_{Wasserstein_{GP}}$ we set a weight of 10, while For KL loss function of UNIT we set a weight of 0.01. We used a mini-batch size of 8 for 2D images with a size of 128×128 and 1 for ones with size 512×512 . In addition, instance normalization was preferred over batch normalization.

4.2.2 Qualitative results on unpaired datasets

Dataset For the unpaired image-to-image translation training, the images were obtained from the Cancer Imaging Archive (TCIA) [24]. Two databases of pathology-free abdominal CT images were used: *CT Pancreas* [118] with 82 ceCT images of abdomens, and *CT Colonography* [124] which contains non-contrast CT images of which 82 healthy subjects were retained for consistency with the first dataset. For both datasets, 72 patients are used for training and 10 for testing. All axial slices were resized from 512×512 to 128×128 pixels to make the training less computationally and memory intensive. Please note that the one modality usually acquired is the ceCT, thus public available CT dataset are hard to gather. We could not find any other public database for CT without contrast other than *CT Colonography*, limiting the possible experiments and qualitative tests to those carried out.

Experiments with state-of-the-art methods First, we tested several existing methods to select the best networks for the generators (G) and the discriminators (D). These include: the original CycleGAN [157] with both U-Net and Res-Net as generators, UNIT [87] and TransGAN [68]. For CycleGAN and UNIT we tested both PatchGAN [157] and Wasserstein Loss [2] as discriminator mechanism. For CycleGAN we tested also U-Net as discriminator as in [145] and the use of attention mechanism in the last layer as in [39].

The use of the only renal region slices was deemed essential for our experiments to obtain anatomically consistent images and our automatic detection of the abdominal region proved effective, removing the need for manual selection. Some qualitative examples about automatic ROI selection is shown in Figure 4.5, in which it is possible to see how both UNIT and CycleGAN show great improvements in the preservation of anatomical structures and general abdominal shape using only the slices of the renal region.

Moreover, in all these tests, we used the PBS [147] strategy for selecting slices at the same relative position. The most interesting combinations that we tested are shown in Figure 4.6, with an example for each domain. CycleGAN showed better results than UNIT, which fails to fully preserve anatomical structures. In order to use attention mechanism at the end of the

U-Net, we had to decrease the depth of the network due to the high demanding memory of this layers. This leads to unrealistic synthetic images (for this reason they are not even presented in Figure 4.6). Instead, the use of U-Net as discriminator creates a gray veil on the image. The original CycleGAN [157] with both generator networks and discriminator mechanisms leads to satisfactory results for the easier task of generating images without contrast. However, only the CycleGAN with Res-Net as the generating network and PatchGAN as the discriminating mechanism produced good results in terms of contrast realness for the task of CT2ceCT, as shown in Figure 4.6. In fact, in this case, the use of a residual network allows for greater anatomical coherence, and the PatchGAN method is the best performing for contrast reproduction, as it is also evident from its use with other methods and other generator networks. For methods such as TransGAN, performances are limited because we had to decrease the number of attention layers used due to reduced computational power. Probably, these results are also caused by the restricted amount of data.

Some experiments using Fréchet Inception Distance (FID) [57] and other proposed measures for unpaired experiments were also performed. The results are shown in Appendix D and some critical issues are discussed.

Experiments with the proposed method Despite the good results shown in the method identified as the best ones in Figure 4.6, in terms of overall shape and contrast intensity, the PBS selection was not sufficient and several anatomical artefacts appeared (see Figure 4.10 and Figure 4.11). Another existing strategy for anatomically-paired selection that we tested was the use of 3D affine registration. Given the high variability between the two domains and the low 3D results using a fixed reference pose shown in Section 3.4.2 with both our proposed STN (see Section 3.4.1 for details) and Simple-Elastix [96], we decided to perform the registration at each iteration between the two selected patients with Simple-Elastix algorithm (details in Appendix B). The anatomical coherence was improved but some important artifacts still appeared, due to the fact that this strategy often failed in building anatomically-paired subjects as can be seen in Figure 4.7 (right), where kidneys are not perfectly aligned after registration. In order to improve results, a 3D registration via landmarks using OpenCV library [11] was tested but resulted to be high-demanding in terms of time and user-interaction. Finally, our proposed selection with SSBR was tested, which resulted in better anatomically-paired selection as shown in Figure 4.9 and consequently reduced the severity of artefacts, as shown in the forth column (top) of Figure 4.10 and Figure 4.11. In fact, in this way the discriminator is able to specialize better in the task. Figure 4.8 shows how using the training of the original SSBR [144] does not guarantee the same score for images of different modalities, conversely from our method with the use of L_{norm} (Equation 4.2). Furthermore, the original training falls into the trivial solution and slices with high anatomical difference appear to have similar scores. The use of the proposed L_{anat} (Equation 4.3) overcomes this weakness.

We then added the L_{ACL} loss function, which significantly improved anatomical coherence, particularly in the binary mask regions. Eventually, we combined in a first moment the use of input as template $I_n A_d$ and in a second moment the binary mask BM . The complete proposed method based on SSBR selection with L_{ACL} , $I_n A_d$ and BM produced high quality synthetic images, without visual artefacts and with realistic contrast intensity according to physicians' evaluation. Some qualitative results are detailed in Figure 4.10 and Figure 4.11.

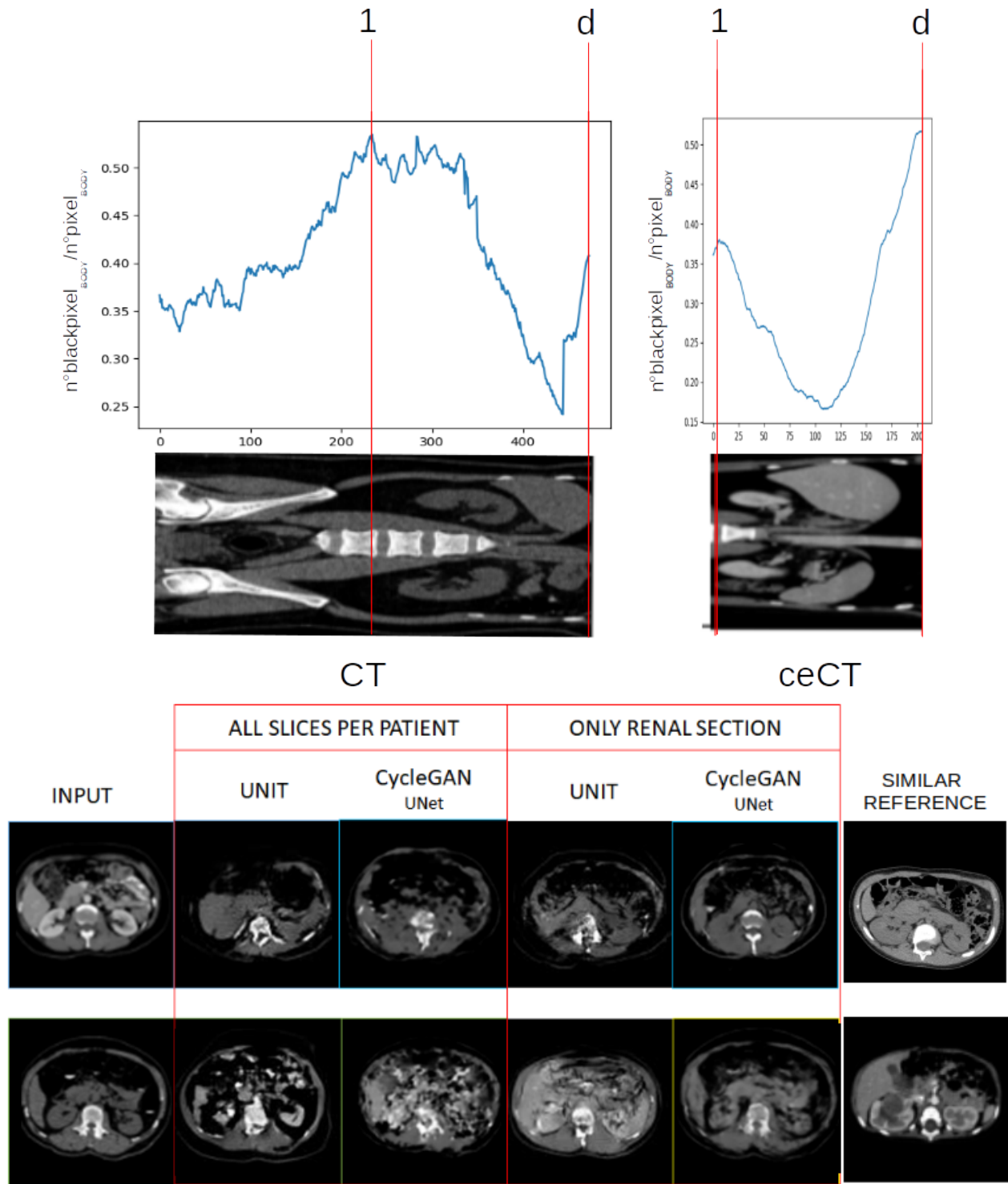


Figure 4.5: Top: example of renal ROI selection using the number of black pixels in relation to the total number of pixels of each individual slice to automatically select the first slice of the lungs and the last slice of the visceral area as the upper and lower reference points respectively. Bottom: comparison of CycleGAN and UNIT trained without and with renal ROI selection (PBS strategy was used). For both methods, we used U-Net as generator network and Patch-GAN as discriminator mechanism. First row: from ceCT to CT. Second row: from CT to ceCT. In the forth and fifth column is shown a general improvement in both the preservation of anatomical structures and the abdominal shape. An idea of how the expected output should look like is provided in the last column.

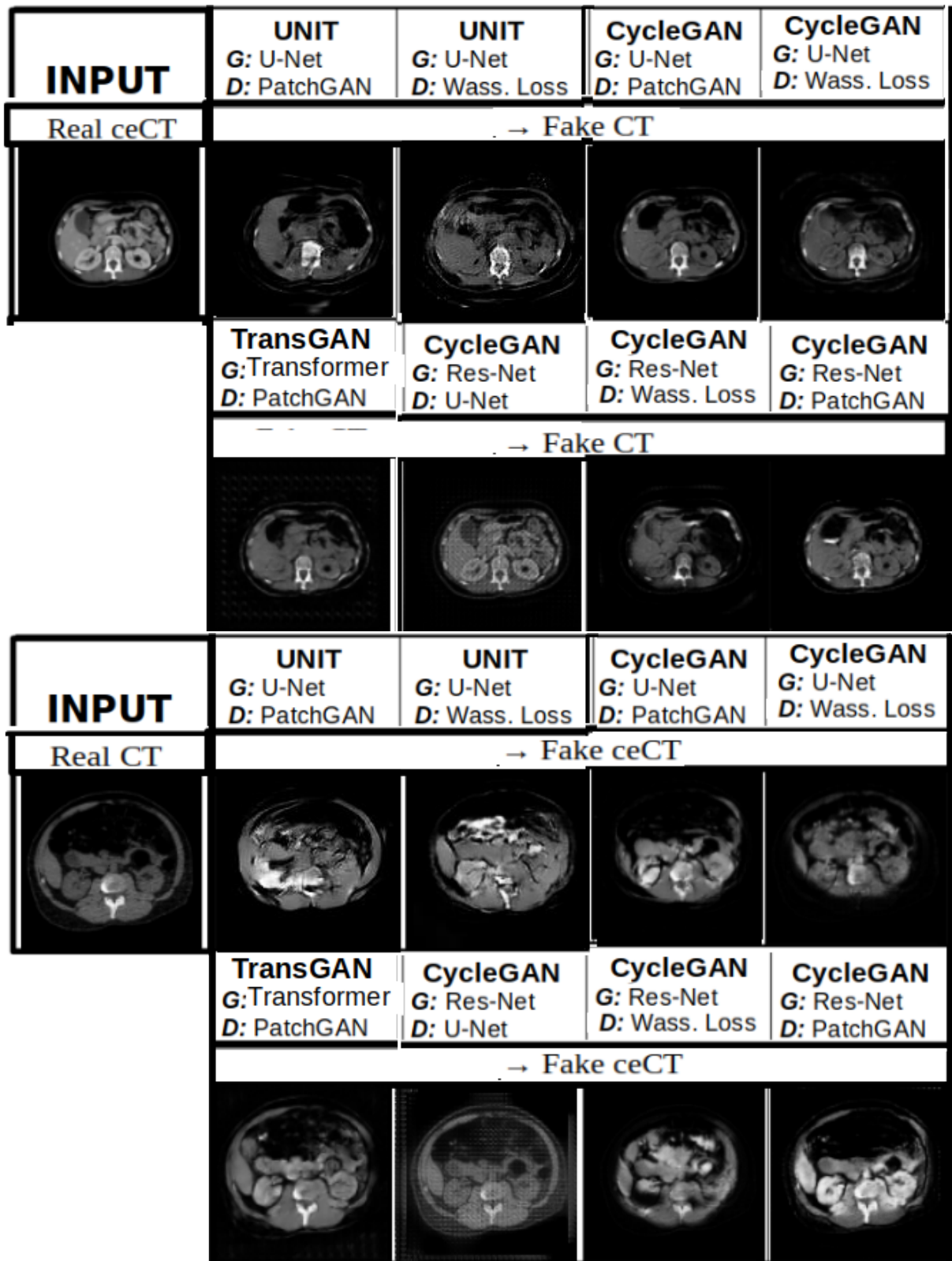


Figure 4.6: Comparison of some state-of-the-art methods on slices of the unpaired test set. Top table: ceCT to CT. Bottom table: CT to ceCT. The slices in all tests are selected with PBS. The input in the other direction gives an idea of what the expected result should look like.

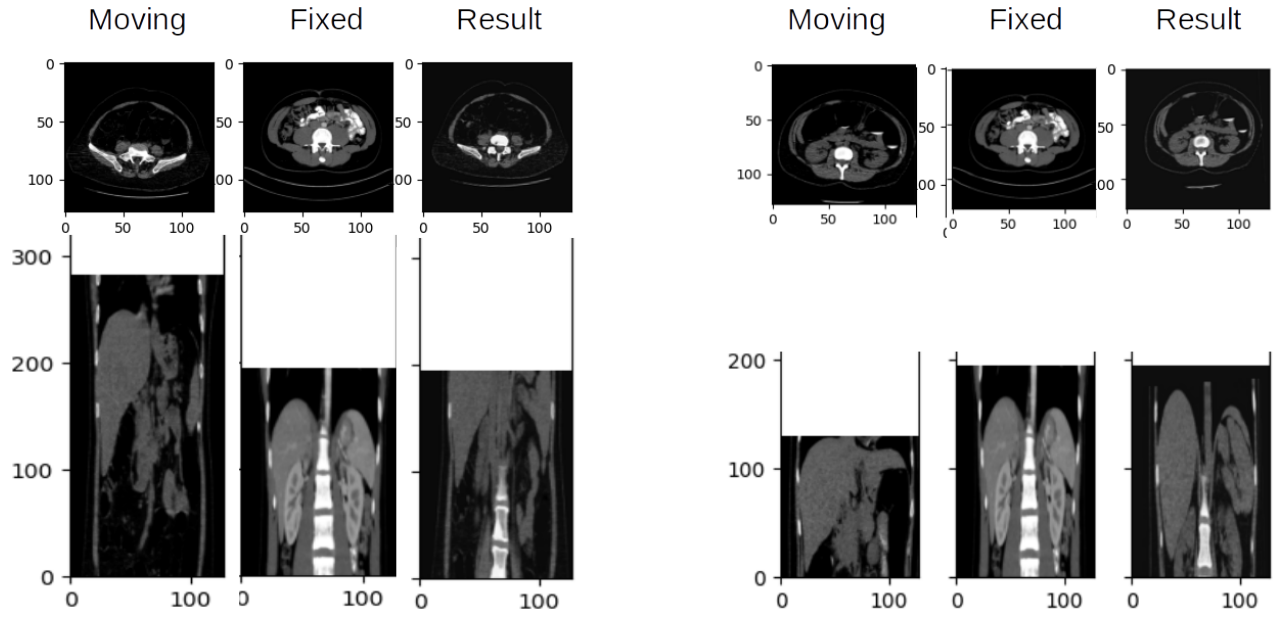


Figure 4.7: Two examples of 3D affine registration using SimpleElastix [96].

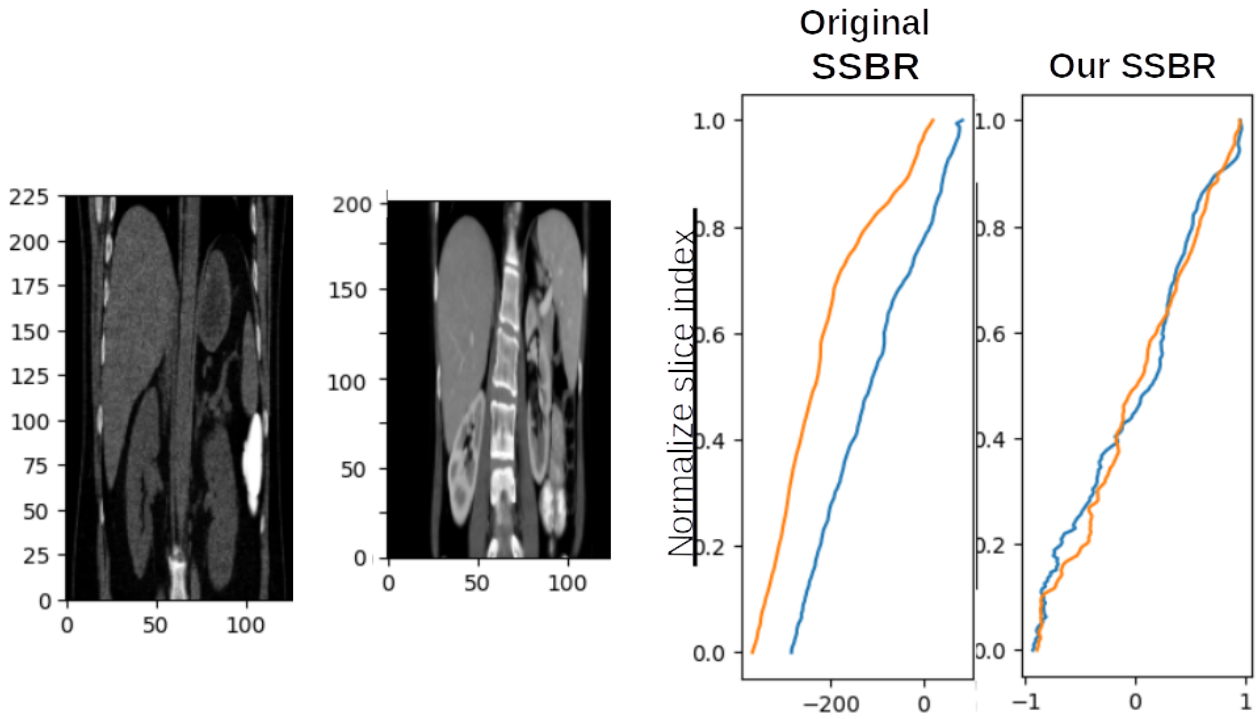


Figure 4.8: Example to show differences in SSBR scores using the original training and our proposed one.

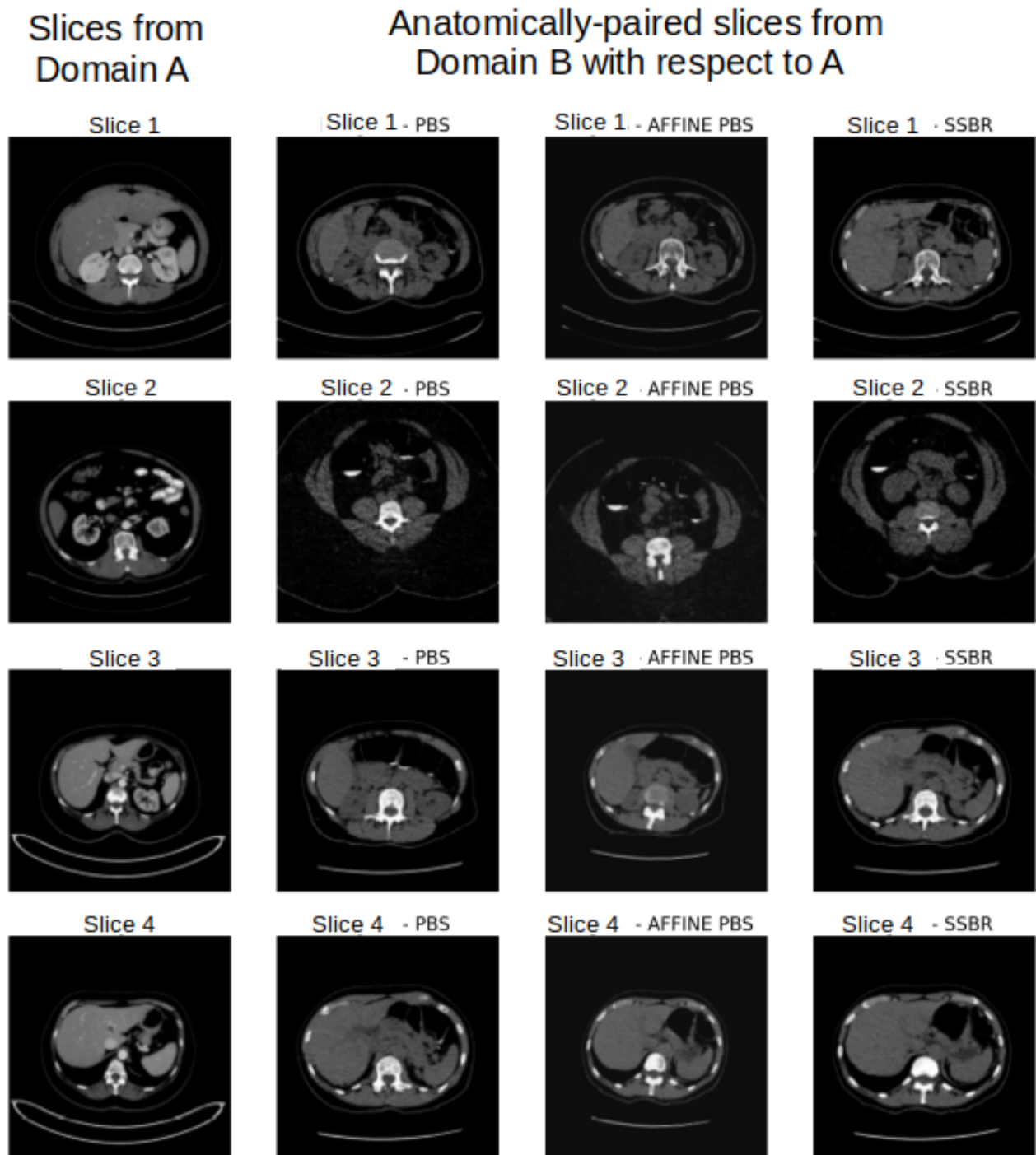


Figure 4.9: Some examples of input selection methods. The anatomically-paired slice B is chosen starting from slice A with a Position-Based Selection (PBS), 3D affine registration+ PBS or our proposed SSBR selection.

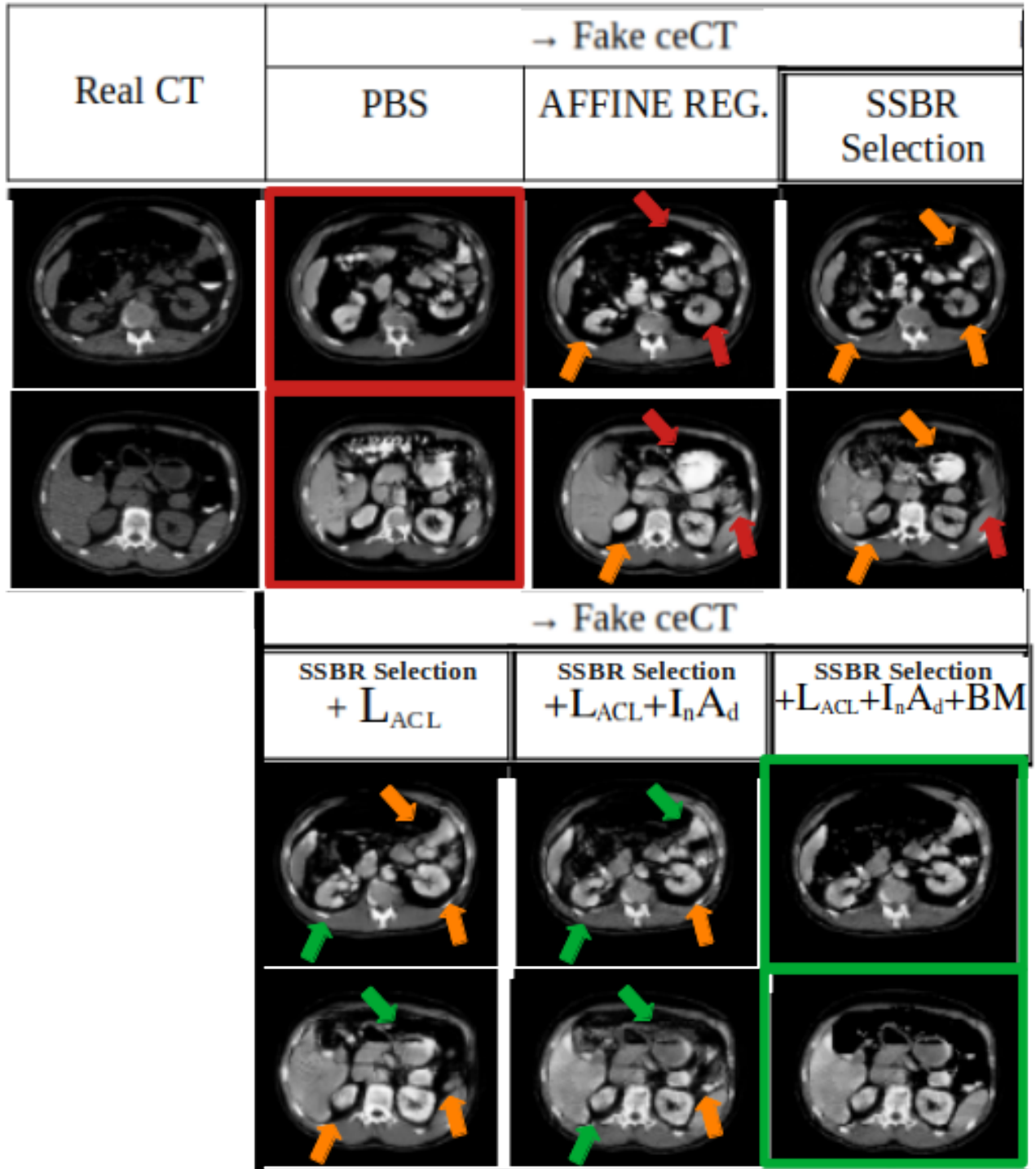


Figure 4.10: Qualitative results on unpaired slices. From CT to ceCT. Tests based on CycleGAN. Three bottom columns: results with our methods, we add each proposition to the SSBR selection. $I_n A_d$ indicates the input addition while BM the use of binary mask. Arrows: high (red), low (orange) and no (green) artefacts.

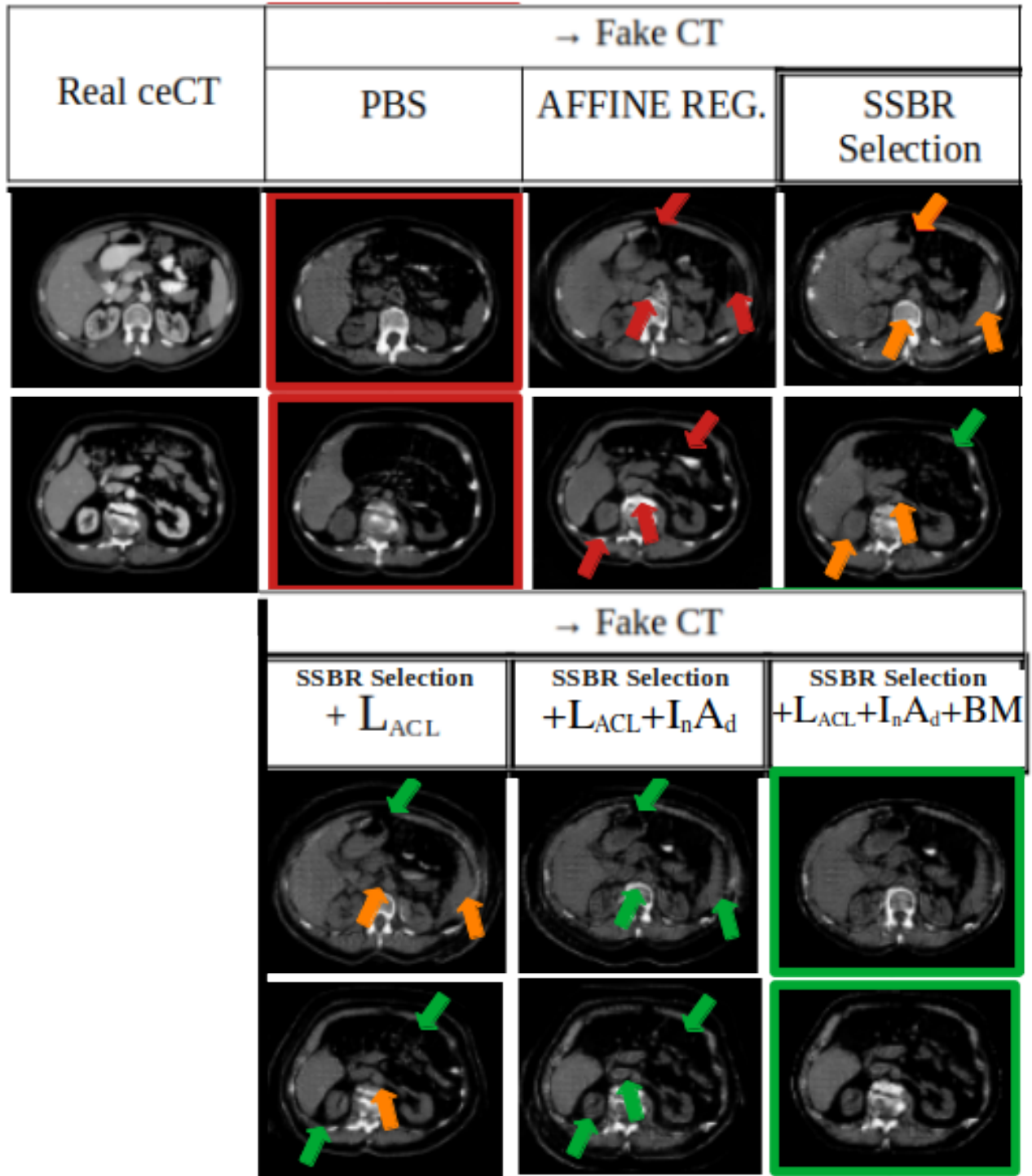


Figure 4.11: Qualitative results on unpaired slices. From ceCT to CT. Tests based on CycleGAN. Three bottom columns: results with our methods, we add each proposition to the SSBR selection. $I_n A_d$ indicates the input addition while BM the use of binary mask. Arrows: high (red), low (orange) and no (green) artefacts.

4.2.3 Quantitative ablation study on paired database

Dataset For quantitative testing we used the only 10 patients of the Necker PRAC database (detailed in Chapter 2) in which paired abdominal ceCT-CT images were available. It is important to note that the small number of patients in this data set prevents achieving satisfactory performance on training generative models. Furthermore, since the ceCT images are acquired a few seconds from the CT, an affine registration using Simple-Elastix [96] slices by slices was performed. No public paired ceCT-CT datasets is available, so this small dataset we gathered is quite rare. Moreover, we want to emphasize that for each subject we have at our disposal about 100 2D slices of the renal ROI and therefore the results refer to a total of about 1000 2D images.

Results The quantitative ablation and comparative study was performed using the presented methods, pre-trained on the unpaired data-sets. The results are presented in Table 4.1 using mean square error (MSE), structure similarity (SSIM) and peak signal to noise ratio (PSNR) between *real* and *fake* images, and training time (TIME). All our contributions improve on the original CycleGAN with PBS [147], at the cost of some additional learning time (note that the inference time remains the same for all methods). Their combination produces the best results for both tasks. The use of affine registration seems to be quantitatively comparable to the use of SSBR for the selection and loss function L_{ACL} , but the network requires a very high computational time in addition to the creation of artefacts, as illustrated in Figures 4.10 and 4.11. Moreover, analyzing these quantitative results, it is important to take into account that the alignment of the paired ceCT-CT may not be perfect.

4.2.4 Blood vessel segmentation using ceCT and CT

To further demonstrate the realness of the images generated by our method, similarly to [23, 119, 125, 158], we compared the performance of a segmentation network when using either a real image and a fake image, or both real images. As discussed at the beginning of the chapter, these authors demonstrate how the use of both CT modalities, combined with *standard iconographic* data augmentation on target structures, alleviates the difficulties caused by the variability in contrast medium diffusion. In the absence of one of the modalities, the use of synthetic images can be exploited. It is also important to point out that another benefit of such an idea lies in the fact that is even more of an effort to produce manual segmentation in CT images compared to ceCT, therefore we can use synthetic CT generated with an image translation method using as reference segmentation that of the real ceCT image. Here, we test this method on pediatric and pathological (renal cancer) subjects for the segmentation of blood vessels, i.e. arteries and veins, in the abdominal-visceral region. These structures present a high heterogeneity in pediatric abdominal ceCT images (see Table 2.1 and Table 2.2), due to the presence of big tumors (with respect to other anatomical structures) which can obstruct the passage of contrast medium. In particular, we use synthetic CT generated from labeled ceCT, in order to train a segmentation network with both real ceCT and synthetic CT.

Dataset For the proposed segmentation application, the synthetic images used were produced using generative methods trained as explained previously but with images at the original size 512×512. We used both the paired dataset presented in the previous section and the complete Necker PRAC dataset for arteries and veins segmentation (detailed in Chapter 2).

Table 4.1: Quantitative study on the 10 patients of Necker database with paired images. Mean square error (MSE), structure similarity (SSIM) and peak signal to noise ratio (PSNR) are shown as mean and standard deviation. TIME is the training time. Tests based on CycleGAN. The last section of rows presents the ablation study, in which each proposition is added to the SSBR selection. I_nA_d = input addition, BM = binary mask.

CycleGAN Method	MSE [10^{-2}] (\downarrow)	SSIM [10^{-1}] (\uparrow)	PSNR (\uparrow)	TIME (\downarrow)
real CT \rightarrow fake ceCT vs real ceCT				
PBS	10.05 (2.89)	5.76 (0.65)	16.14 (1.15)	3h 2m
AFFINE REG.	8.16 (1.80)	6.36 (0.57)	16.99 (0.87)	16h 33m
SSBR selection	9.07 (2.39)	5.99 (0.71)	16.56 (1.07)	7h 5m
+ L_{ACL}	8.55 (2.28)	6.19 (0.69)	16.82 (1.07)	7h 49m
+ BM	8.42 (2.46)	6.24 (0.73)	16.91 (1.17)	7h 5m
+ I_nA_d	6.79 (2.85)	6.60 (0.74)	17.97 (1.54)	7h 14m
+ $L_{ACL} + BM$	8.19 (2.32)	6.36 (0.72)	17.02 (1.14)	7h 49m
+ $L_{ACL} + I_nA_d$	6.41 (1.97)	6.67 (0.63)	18.11 (1.22)	7h 55m
+ $L_{ACL} + I_nA_d + BM$	6.37 (2.01)	6.81 (0.62)	18.14 (1.23)	7h 55m
real ceCT \rightarrow fake CT vs real CT				
PBS	8.26 (1.97)	5.36 (0.28)	16.96 (1.04)	3h 2m
AFFINE REG.	4.72 (0.95)	6.77 (0.37)	19.36 (0.93)	16h 33m
SSBR selection	7.15 (2.16)	5.68 (0.52)	17.64 (1.26)	7h 5m
+ L_{ACL}	5.87 (1.73)	6.08 (0.22)	18.47 (1.12)	7h 49m
+ BM	6.07 (1.28)	6.61 (0.65)	18.28 (0.99)	7h 5m
+ I_nA_d	6.16 (1.15)	5.87 (0.23)	18.18 (0.79)	7h 14m
+ $L_{ACL} + BM$	5.08 (0.85)	6.87 (0.52)	19.02 (0.74)	7h 49m
+ $L_{ACL} + I_nA_d$	4.24 (0.86)	6.80 (0.37)	19.83 (0.92)	7h 55m
+ $L_{ACL} + I_nA_d + BM$	4.05 (0.83)	7.23 (0.53)	20.03 (0.92)	7h 55m

Reference segmentations of arteries and veins were manually performed by medical experts of Necker also on the paired CT images due to possible misalignment. Eventually, in order to test the method with all the labeled structures in the complete Necker PRAC dataset, we test it also for ureters segmentation and for kidneys and tumor segmentation (see Chapter 2).

Segmentation performances Given the restricted paired dataset, all tests were done with the Leave-One-Patient-Out (L-O-P-O) method using the 3D nnU-Net [61] framework, detailed in Section 3.1.2. In order to be consistent, the weights initialisation and the research space ($N_{patches} \times N_{iterations}$) are fixed for all the nnU-Net trainings. Results (using evaluation measures of Appendix A) show that replacing a real CT modality with a synthetic one produced with CycleGAN and the PBS method, as in [119, 125], is not sufficient to achieve performances as good as when using both real modalities. By contrast, the synthetic CT images produced by our method achieve the highest Dice score and the lowest Hausdorff distance, with the best combination of precision and recall. Moreover, these results achieve almost same performances of using both real modalities. This is even more evident for the more heterogeneous cases, particularly for the veins. Quantitative results are shown in Table 4.2, while some qualitative results are illustrated in Figure 4.12. We can see that the use of both modalities leads the network to focus less on the HU value, and the use of our fake images produces results very similar to those with the real ones. These improvements result in better segmentation in areas with strong heterogeneity and less confusion with structures with similar contrast.

For the sake of completeness, even if not present in any existing methods, experiments using 3D affine registration and PBS were also performed, showing comparable results of using just PBS (arteries $DS = 70.52\%$ ($sd=7.74$), veins $DS = 42.77\%$ (19.21)).

Table 4.2: Segmentation performance on **real ceCT** of 10 patients (and then on the only 5 more heterogeneous cases) using L-O-P-O methods. Dice score (DS), precision (PR), recall (RC) and 95th percentile of the Hausdorff distance (95HD) are given in mean and standard deviation. All tests were done using 3D nnU-Net [61] with intensity (except if indicated with “no DAI”) and geometric data augmentation (see Appendix B). The first row for both sections is our goal.

INPUT Database	Structure	DS [100%] (\uparrow)	PR [100%] (\uparrow)	RC [100%] (\uparrow)	95HD [mm] (\downarrow)
on 10 patients					
real ceCT and real CT	Arteries	74.61 (5.89)	85.22 (8.32)	69.06 (8.15)	15.39 (5.72)
	Veins	45.62 (13.72)	60.61 (19.53)	38.68 (14.83)	31.47 (16.53)
real ceCT <i>no DAI</i>	Arteries	63.75 (11.18)	80.33 (10.99)	53.88 (12.48)	23.43 (8.18)
	Veins	21.18 (19.70)	64.04 (34.08)	15.45 (16.04)	42.14 (23.79)
real ceCT	Arteries	73.01 (6.57)	81.08 (8.70)	67.19 (8.43)	15.80 (7.01)
	Veins	40.58 (23.50)	55.94 (31.39)	33.72 (26.61)	40.65 (30.90)
real ceCT and fake _{PBS} CT	Arteries	69.59 (8.89)	79.54 (10.85)	63.47 (12.59)	18.08 (8.21)
	Veins	44.40 (22.75)	58.44 (21.78)	38.38 (23.20)	39.31 (16.79)
real ceCT and fake _{Ours} CT	Arteries	72.33 (7.41)	77.29 (10.32)	68.63 (8.88)	15.48 (6.38)
	Veins	44.49 (22.50)	54.98 (26.74)	40.28 (22.69)	38.90 (32.76)
on the 5 more heterogeneous cases					
real ceCT and real CT	Arteries	75.01 (5.82)	85.17 (4.37)	67.50 (8.57)	12.79 (6.04)
	Veins	40.87 (14.73)	56.93 (18.63)	32.62 (13.05)	31.16 (10.76)
real ceCT <i>no DAI</i>	Arteries	66.59 (8.31)	86.89 (5.70)	54.83 (10.29)	23.34 (9.14)
	Veins	14.66 (17.05)	71.31 (39.90)	8.89 (10.98)	50.35 (29.50)
real ceCT	Arteries	72.94 (6.30)	84.37 (3.80)	64.89 (9.71)	13.49 (5.14)
	Veins	28.28 (19.84)	51.97 (38.06)	17.50 (18.41)	35.57 (14.33)
real ceCT and fake _{PBS} CT	Arteries	70.77 (9.18)	84.41 (5.96)	63.00 (15.51)	13.83 (5.95)
	Veins	33.47 (26.92)	45.48 (34.33)	27.73 (23.78)	37.73 (23.42)
real ceCT and fake _{Ours} CT	Arteries	73.18 (7.51)	80.58 (4.59)	67.63 (11.25)	12.73 (4.10)
	Veins	40.57 (20.25)	62.01 (13.31)	31.96 (18.91)	32.83 (13.84)

Other quantitative results for the entire ceCT Necker dataset (no CT images available) for the segmentation of arteries and veins can be found in Table 4.3. Here, we trained the nnU-Net [61] for segmenting arteries and veins using 51 early-injected ceCT pediatric patients. The remaining 15 subjects are used as test set. The advantages of using both real ceCT and the synthetic CT images generated by our method instead of PBS are even more visible. This is due to the fact that we consider a larger dataset with many more heterogeneous images, so using both ceCT and CT images with high anatomical consistency between the images and the reference segmentation, produces the best result. However, since we do not have the paired real CT images, we cannot verify what the expected results are. Eventually, on the same way, we test the method also for ureters segmentation (Table 4.4) and for kidney and renal tumor segmentation (Table 4.5). The use of synthetic CT images generated by our method leads to the best performances. Nevertheless, here the benefits of the combined use of real ceCT and synthetic CT are less visible than in arteries and veins. This may be caused by the few data in

the ureters database and the lower heterogeneity of the kidneys and tumors. It is interesting to note how in kidney and tumor segmentation, the exploitation of images generated by the PBS method produces segmentation results much worse than using just real ceCT. This is probably due to an anatomical inconsistency between CT input and reference segmentation.

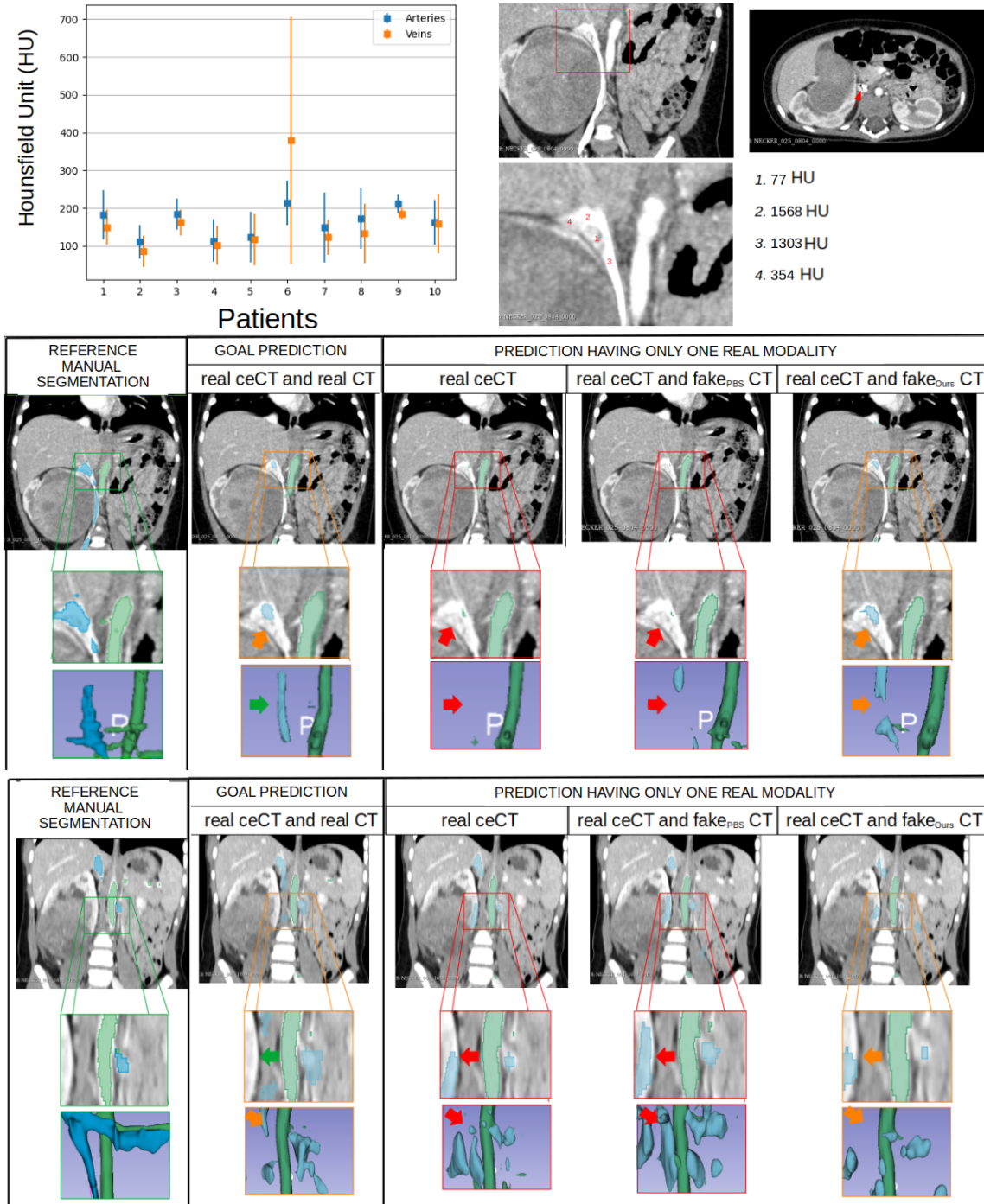


Figure 4.12: Top left: box plot representing mean and standard deviation of HU for abdominal blood vessels for the 10 paired patients. Top right: Particular on patient 6 where the tumor blocks the injection of contrast and it causes high heterogeneity on cava vein. Bottom: Segmentation results of the most heterogeneous patient (top, patient 6) and the least heterogeneous one (bottom, patient 9). Arteries are displayed in green, and veins in blue. Arrows: strong (red), light (orange) and no (green) error.

Table 4.3: Segmentation performance on **real ceCT** of 15 patients of the **test set of arteries and veins ceCT Necker dataset**, composed of 65 patients (we used 43 for training and 7 for validation). Dice score (DS), precision (PR), recall (RC) and 95th percentile of the Hausdorff distance (95HD) are given (mean and standard deviation). All tests were done using 3D nnU-Net [61] with intensity and geometric data augmentation.

INPUT Database	Structure	DS [100%] (\uparrow)	PR [100%] (\uparrow)	RC [100%] (\uparrow)	95HD [mm] (\downarrow)
on 15 patients					
real ceCT	Arteries	63.45 (5.67)	71.73 (9.99)	57.87 (7.31)	17.46 (9.65)
	Veins	42.64 (20.12)	76.67 (13.17)	31.84 (17.12)	23.55 (17.00)
real ceCT and fake _{PBS} CT	Arteries	65.60 (4.45)	73.04 (10.83)	60.91 (7.12)	15.59 (8.47)
	Veins	45.77 (18.67)	73.14 (14.88)	35.37 (17.87)	21.25 (20.05)
real ceCT and fake _{Ours} CT	Arteries	70.01 (3.99)	76.29 (8.23)	65.77 (7.73)	13.47 (10.09)
	Veins	56.55 (20.20)	81.53 (8.91)	46.98 (22.38)	20.93 (22.96)
on the 5 more heterogeneous cases					
real ceCT	Arteries	63.23 (4.24)	74.86 (7.53)	54.99 (4.55)	15.54 (6.08)
	Veins	27.43 (20.62)	66.64 (15.59)	19.90 (17.58)	24.90 (8.42)
real ceCT and fake _{PBS} CT	Arteries	64.97 (1.12)	76.68 (11.92)	57.61 (5.97)	15.49 (5.26)
	Veins	33.16 (18.83)	62.77 (18.28)	24.18 (16.09)	20.91 (7.55)
real ceCT and fake _{Ours} CT	Arteries	70.15 (3.52)	80.40 (9.97)	62.89 (4.71)	12.15 (6.65)
	Veins	37.00 (16.11)	77.58 (11.78)	26.01 (14.02)	21.71 (6.33)

Table 4.4: Segmentation performance on **real ceCT** of 5 patients of the **test set of ureters ceCT Necker dataset**, composed of 16 patients (we used 10 for training and 1 for validation). Dice score (DS), precision (PR), recall (RC) and 95th percentile of the Hausdorff distance (95HD) are given (mean and standard deviation). All tests were done using 3D nnU-Net [61] framework with intensity and geometric data augmentation.

INPUT Database	Structure	DS [100%] (\uparrow)	PR [100%] (\uparrow)	RC [100%] (\uparrow)	95HD [mm] (\downarrow)
on 5 patients					
real ceCT	Ureters	54.43 (24.82)	78.95 (13.19)	49.28 (29.35)	19.46 (24.76)
real ceCT and fake _{PBS} CT	Ureters	54.02 (28.61)	73.31 (8.01)	51.73 (31.28)	17.58 (26.39)
real ceCT and fake _{Ours} CT	Ureters	57.57 (21.32)	73.34 (6.93)	53.23 (26.14)	13.99 (25.34)

Table 4.5: Segmentation performance on **real ceCT** of 15 patients of the **test set of kidneys and renal tumor ceCT Necker dataset**, composed of 80 patients (we used 58 for training and 7 for validation). Dice score (DS), precision (PR), recall (RC) and 95th percentile of the Hausdorff distance (95HD) in mean and standard deviation. All tests were done using 3D nnU-Net [61] with intensity and geometric data augmentation.

INPUT Database	Structure	DS [100%] (\uparrow)	PR [100%] (\uparrow)	RC [100%] (\uparrow)	95HD [mm] (\downarrow)
on 15 patients					
real ceCT	Kidneys	90.15 (3.57)	91.22 (3.31)	89.80 (9.11)	4.59 (3.58)
	Tumors	86.92 (10.49)	94.00 (5.80)	82.35 (17.06)	8.20 (6.39)
real ceCT and fake _{PBS} CT	Kidneys	87.53 (9.56)	90.03 (3.35)	86.36 (13.75)	4.77 (3.18)
	Tumors	82.17 (25.12)	88.46 (19.20)	77.13 (27.65)	10.70 (12.08)
real ceCT and fake _{Ours} CT	Kidneys	90.84 (3.56)	89.46 (4.20)	92.38 (4.17)	4.62 (4.39)
	Tumors	88.39 (11.71)	92.17 (9.40)	86.80 (14.96)	6.46 (6.54)

4.3 Conclusion

In this chapter we presented an extension of CycleGAN via the use of a Self-Supervised Body Regressor to: (i) better select anatomically-paired slices; (ii) anatomically constrain the generator to produce a slice describing the same anatomical content as the input. This method can be used independently of the choice of generating network and discrimination mechanism. It is important to highlight that the world of GAN-based methods is progressing rapidly, and more and more advanced and high-performance methods are being proposed. We have made choices both because of technical limitations and because of what was discussed at the beginning of the chapter on carbon emissions. In fact, for a fairer comparison we would have had to try more methods and see if proposed modifications lead to significant changes as well. However, we believe that the methods examined are sufficient to cover much of the state of the art, especially if we focus on the world of GANs applied in the medical domain. In fact, in this field, as we can see from the articles mentioned here that date from the last two years, the methods adopted are those that we analyze in this chapter. This is probably due to the large size and quality of medical images (high memory required) as well as the restricted databases, especially in the pediatric field (difficulty of convergence or generalization for the networks).

The method we propose is designed from the difficulties associated with the translation of unpaired abdominal ceCT-CT images but it is applicable on other medical translation tasks. However, we believe that: (i) the higher spatial consistency in other anatomical areas (brain or lungs), may make the use of the PBS method sufficient for the selection of anatomically-paired slices; the substructures differences between different acquisition modalities (e.g. MR and CT or PET and CT) ease the discrimination by the critic mechanism and the use of anatomically-paired slices may not be necessary, as demonstrated by some work in the literature. Finally, all the difficulties presented would be overcome with the use of 3D networks that examine the entire patient volume. Nevertheless, this scenario would require GPUs with very high VRAM, which I do not believe are currently available, as well as an infinite training time.

Refocusing on the translation of abdominal ceCT-CT examined in this chapter, we showed significant improvements in the generated images compared to existing methods. The difficulty of finding paired images, especially in the pediatric field, is the central point of this study which at the same time leads to a lack in quantitative evaluation. In fact, unpaired images do not yet allow reliable quantitative assessments (see Appendix D) and visual assessment can be difficult. The joint work with medical experts allows to have more solid qualitative evaluations, understanding which images have still significant artifacts.

To further validate our method, we demonstrated that the synthesized images can be used to guide a segmentation method by compensating, without loss of performance, for the absence of the complementary real acquisition modality. However, the improvement given by the use of both modalities is not visible for the segmentation of all the structures. The use of a strong iconographic data augmentation sometimes seems to be sufficient.

Before concluding, we want to point out that the segmentation results showed in this chapter for arteries, veins and ureters are still unsatisfactory from a clinical point of view. For this reason segmentation methods for renal tubular structures are specifically examined in the next chapter.

Chapter 5

Segmentation of tubular structures on pediatric abdominal-visceral ceCT scanners with renal tumor

In order to achieve our final goal of a complete 3D model of a patient affected by renal tumor, we now address the automatic segmentation of arteries, veins and collecting system (i.e. ureters). These structures raise several challenges, such as (i) elongated shape (few pixels in a 2D cross section), (ii) intra- and inter-patient contrast heterogeneity, and (iii) intra-scale changes. Some authors [51, 85, 128] try to solve these issues on adults ceCT images without great results or limited to specific acquisition CT modalities. The difficulties of segmenting such tubular structures are also increased in pediatric subjects, due to:

- *Inter-anatomy variation.* Pediatric databases include subjects with ages from 1 day to 16 years, thus the size and position of the vessels vary widely among subjects. Moreover in case of pathology (e.g. renal tumor) the shape and direction of the vessels also vary (difficulties to choose a 2D cross section to work on).
- *Small volume to background ratio.* In addition to the problem of intra-scale changes, given by the difference in vessel diameters also found in adults, in pediatric subjects such structures are very small (see anatomical information in Chapter 2). Even using patches, these structures will still result in a small number of foreground voxels compared to the background ones, as shown in Figure 5.1.
- *Small available labeled dataset.* Pediatric databases are limited in number of images as discussed in Chapter 3.

We propose, for the first time, an automatic segmentation approach of renal tubular structures for *pediatric* and *pathological* patients. Our method is based on CNN, merging the best of the state-of-the-art methods and adding a new loss function built on *vesselness* [38, 80, 84], determined from the eigenvalues of the Hessian matrix of segmentation masks. The main features of our propositions are:

- Comparing the eigenvalues of the Hessian matrix of the predicted and reference segmentations forces the network to learn the morphology of the target structure. These eigenvalues have specific patterns depending on shape: they are different from zero in

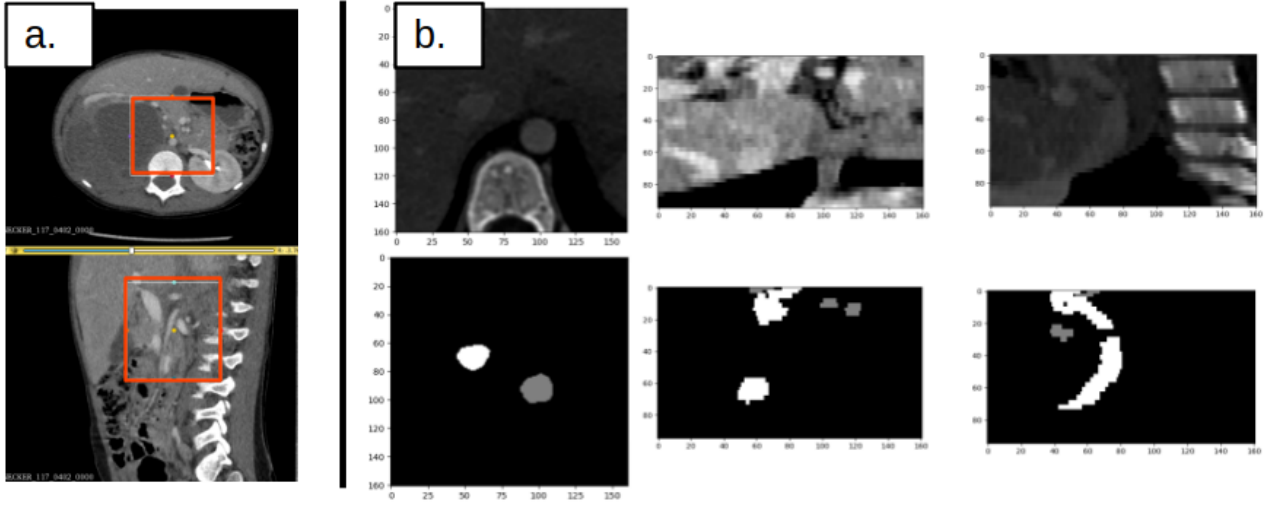


Figure 5.1: Example of patches of size $96 \times 160 \times 160$ extracted from the input volume with voxel size $0.9 \times 0.46 \times 0.46 \text{ mm}^3$ ($Z \times X \times Y$). a. Relation of patch with respect to the input volume (top: axial view, bottom: sagittal view). b. Example of patches with elongated structures in (from left to right) axial view, coronal view and sagittal view (top: input images, bottom: reference segmentation). The patches in a. and b. are not corresponding.

the case of elongated structures (with one preferred direction, one eigenvalue is much smaller in magnitude than the other two) or flattened structures (two preferred directions) such as vessels (which are elongated but can also be compressed in some regions by other organs or tumors because of the lower rigidity, as for the veins [132]).

- Instead of being used on the input image, the vesselness function is adapted to be used on segmentation masks for the first time. For this reason some modifications to this function are developed.
- The different sizes of the vessels are considered using Gaussian filters with multiple standard deviation. The combination of this technique with deep supervision also makes it possible to operate on slices or patches with vessels of different diameters, without having to handle any additional hyperparameter.
- The above considerations are modeled as loss functions and combined with voxel-wise ones such as Dice score and cross-entropy. This approach improves the performance of elongated tubular structures segmentation in comparison with the use of voxel-wise loss functions alone.
- For computation efficiency, it is possible to limit the time-consuming calculation of eigenvalues to the only voxels of the target structure or to a dilation of it.

This chapter is divided into three main contributions. First, a complete assessment on state-of-the-art methods for the segmentation of renal tubular structures on ceCT images on adults is presented in Section 5.1. Secondly a comparison of these methods on adults is performed on pathological and pediatric ceCT images in Section 5.2.1. To the best of our knowledge, both assessment and comparison on this specific case are novel. Eventually, the best techniques identified are merged with a proposed oversampling method and improved with

the use of the proposed tubular structures loss function based on vesselness. A comprehensive study of this is detailed in Section 5.2.2.

The rest of the chapter is organized as follow. In Section 5.3 the images selected from the Necker PRAC database are presented as well as the implementation of the experiments performed. Then, in Section 5.4 the results on both the state-of-the-art comparison and the proposed method are shown and discussed. Finally, in Section 5.5 our conclusion are drawn.

5.1 Assessment of State-Of-The-Art methods of renal tubular structures segmentation on ceCT images

Most of the studies on vessel segmentation that are applied to 3D contrast-enhanced imaging modalities (for both CT and MRI) on adults are extensively described by Lesage *et al.* [84] for non-machine learning-based methods (which we will refer to as “rule-based” for simplicity) and by Moccia *et al.* [102] for machine learning methods, in particular deep learning. In these reviews, it appears that a popular approach is the use of second-order derivative information, captured via Hessian-based filters, to characterize the local image geometry. These techniques can be summarized under the name “vesselness filters” (as done by Lamy *et al.* in [79, 80]) and because of their importance, as well as inspiration for our proposed method, the first part of this section is addressed to them. It is important to emphasize that we mainly focus on methods specifically presented for renal tubular structures segmentation on ceCT, although some relevant works from other domains are also introduced.

5.1.1 Vesselness filters

Vessels in contrast-enhanced medical images are characterized by hyper-intensity and specific geometry features. Therefore, they can be seen as a bright tubular structure on a dark background and consequently for a given function that analyzes these properties, the voxels of the vessels will have a higher score, namely *vesselness*. A fair amount of work was dedicated to the proposition of such a function (see [79, 80, 84] for reviews), and most of them arising from the analysis by Lorenz *et al.* [89]. This states that a voxel can be considered belonging to a vessel if the Hessian matrix computed at this voxel, has a small eigenvalue of either sign and the other two eigenvalues are large and negative. Let (x, y, z) be the space coordinates and $f(x, y, z)$ be the intensity value of the image (defined by a function f), the Hessian matrix H is defined as:

$$H(f) = \begin{bmatrix} h_{xx} & h_{xy} & h_{xz} \\ h_{yx} & h_{yy} & h_{yz} \\ h_{zx} & h_{zy} & h_{zz} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial xy} & \frac{\partial^2 f}{\partial xz} \\ \frac{\partial^2 f}{\partial yx} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial yz} \\ \frac{\partial^2 f}{\partial zx} & \frac{\partial^2 f}{\partial zy} & \frac{\partial^2 f}{\partial z^2} \end{bmatrix} \quad (5.1)$$

and calculated at each space coordinate. Let W_1 , W_2 and W_3 be the three normalized eigenvectors of $H(f)$, associated with the eigenvalues λ_1 , λ_2 and λ_3 , respectively, with $|\lambda_1| < |\lambda_2| < |\lambda_3|$. The vesselness is characterized by:

$$\begin{aligned} |\lambda_1| &\approx 0 \text{ (namely small)} \\ \lambda_2 &\ll 0 \\ \lambda_3 &\ll 0 \end{aligned} \quad (5.2)$$

In order to deal with the non-continuity of the digital medical images and with the high sensitivity to noise of second order derivatives, the Hessian matrix is computed on the image convolved with a Gaussian kernel g_σ . Moreover, in presence of a vessel, W_1 associated to λ_1 corresponds to the direction of the putative vessel, while W_2 and W_3 form a basis of the vessel cross section where $|\lambda_2|$ and $|\lambda_3|$ represent the sizes of the cross section. The most popular function to score the so-called vesselness is the one proposed by Frangi *et al.* [38]. Aiming to build a method suited for medical images, they developed a filter F based on three measures and three parameters:

$$F = (1 - \exp(\frac{-R_a^2}{2\alpha^2})) \exp(\frac{-R_b^2}{2\beta^2})(1 - \exp(\frac{-S^2}{2\gamma^2})) \quad (5.3)$$

if $\lambda_2, \lambda_3 < 0$ and $F = 0$ otherwise, with:

$$\begin{aligned} R_b &= |\lambda_1| / \sqrt{|\lambda_2 \lambda_3|} \\ R_a &= |\lambda_2| / |\lambda_3| \\ S &= \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2} \end{aligned} \quad (5.4)$$

in which R_b discriminates blobs, R_a distinguishes plate and line structures, S measures the norm of the Hessian matrix to avoid enhancing low contrast structures, and α, β and γ control the importance of each measure. Moreover, to cope with different vessel sizes, the vesselness measure is analyzed at different Gaussian kernels and then the maximum response, i.e. the maximum F , is kept. As stated in [84], Hessian-based filters may suffer from sensitivity to local deformations (bifurcations, thrombus or flattened vessels). Moreover, a parameters search is required for both the scale-space parameters (e.g. σ of g) and the intrinsic parameters of the methods (e.g. α, β and γ in Frangi's vesselness). In addition, extraction of large vessels requires large standard deviations of g_σ that can result in perturbation in the response due to other bright structures in the immediate vicinity of the target vessel. These limitations are confirmed by the study presented in [80] where acceptable segmentation results using vesselness filters are obtained only for synthetic images that present no structures other than vessels. However, they can be useful as a first step for segmenting ceCT images, in particular for intra-organs vessels where the neighborhood is more homogeneous (further details in Section 5.1.2). Some interesting alternatives to multiscale Hessian-based filters have been developed, such as the optimally oriented flux (OOF) [83], the ranking of the orientation responses of path operators (RORPO) [99] and the work in [5] where the Frangi's vesselness is applied on the local Jacobian of the gradient vector flow (GVF) field. However, OOF and RORPO are also analyzed in [80] and no improvement in performance over Hessian-based filters was observed, while the method presented in [5] has high computational cost, also in GPU, due to GVF diffusion calculation.

In literature, besides segmentation purpose, vesselness functions have also been used as cost function in lung vessel-tree registration algorithm for ceCT images [16, 22, 33]. Small vessels give almost no contribution to intensity-based similarity metric, thus to further improve the registration accuracy, the sum of squared vesselness measure difference (SSVMD) is employed as geometric-feature similarity metric. In fact, the lung region-of-interest (ROI) in ceCTs perfectly fits the idea of bright tubular structures in dark background (i.e. air is black in ceCTs), overcoming some weakness of Hessian-based filters previously exposed. With the same goal, very recently Wange *et al.* [136] proposed the use of vesselness as a loss function

in a CNN for the first time. In particular they use a normalized vesselness Jerman filter [66], that is more robust to weakly contrasted regions than Frangi's one.

5.1.2 Rule-based methods for tubular structures segmentation

Works that do not use neural networks [84] usually rely on the fact that vessels are contrasted with respect to other structures, i.e. a bright (or dark) structure in a dark (resp. bright) background, as the previously presented *vesselness filters*. On abdominal-visceral ceCT images, this case is restricted to early arterial phase acquisition for having high contrasted arteries or to late delayed phase acquisition for having high contrasted excretory pathways (i.e. ureters). Examples of each phase acquisition time in relation to contrast agent injection in CT are displayed in the Introduction in Figure 1.4. Nevertheless, as explained in the Introduction, due to the very rapid occurrence of the arterial phase (in particular on children) or due to medical choice in order to have both arteries and veins with higher intensity, very often ceCT images are acquired in a late arterial phase, also known as arteriovenous phase [54]. Here, both types of vessels are contrasted but the presence of less contrast medium in each one results in lower intensity (and thus less difference with other structures) as well as greater heterogeneity (which may be accentuated by the presence of tumor or thrombus). Some examples are shown in Figure 5.2.

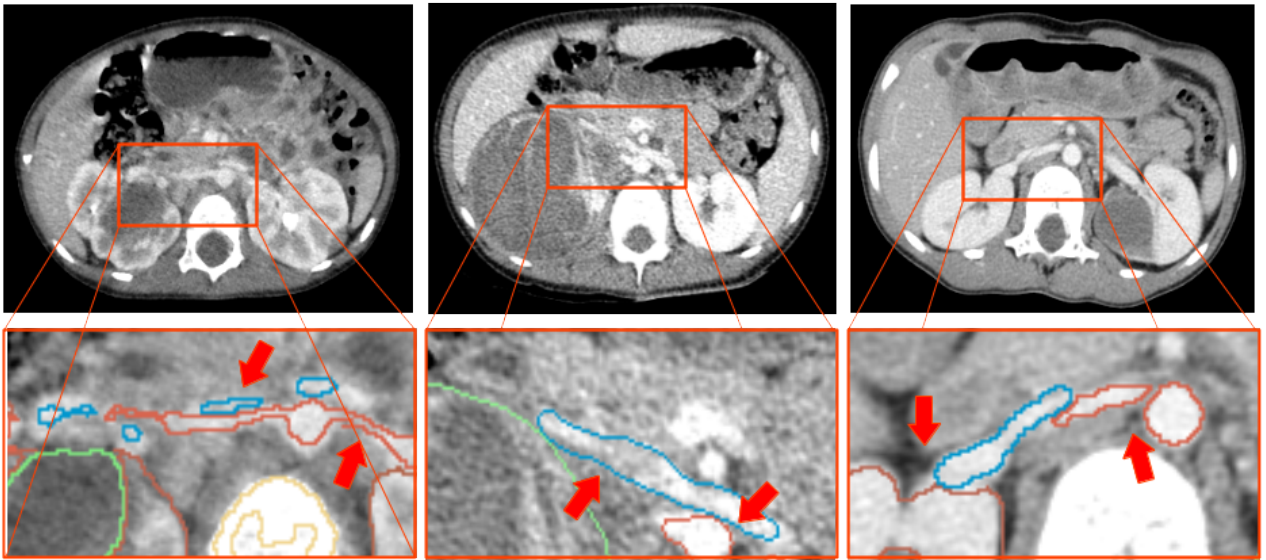


Figure 5.2: Some examples of arteriovenous phase ceCT images from the pediatric and pathological dataset of Necker hospital. From left to right: low contrast in both veins (left arrow) and arteries (right arrow); heterogeneity zone in veins (left arrow) caused by the tumor, while high intensity on arteries (right arrow); renal parenchyma (left arrow) and blood vessels (right arrow) shows similar intensities. Color code for contours: kidneys in brown, tumor in green, arteries in red, veins in blue.

To tackle this problem, Bugajska *et al.* [13] present a semi-automatic approach based on three steps for the segmentation of renal tubular structures in ceCT composed of: (i) a first binarization and erosion, in which the threshold is derived from a ROI (derived by points defined by the user); (ii) a subsequent Locally Adaptive Region Growing (LARG) technique to deal with the lack of homogeneity of voxel intensity values due to an improper contrast propagation; (iii) a final level set method (LSM) which uses the result of LARG

as initialization. Results on 10 patients show that the first two steps are already sufficient to obtain good results in segmentation of the main and larger abdominal-visceral arteries, while only after the third step the small renal vessels reach 80% of the Dice score. However, the method still works better on ceCT with *non-late* arterial phase, otherwise there is a high demand of user-interaction. Yet at the same time, the method does not allow for the segmentation of veins, in which the contrast is still too low. A similar approach is used in [54], in which the authors point out the difficulties of dealing with a combined arteriovenous phase in ceCT images, and for this reason the main threshold is determined by analyzing different points defined by the user. Then they focus on intra-renal vessels segmentation using a LARG on a manually-delineated ROI around each kidney in which a vessel enhancement through multiscale vesselness Frangi filter [38] was first applied. Among unsupervised methods not based on deep learning, it is also important to mention Tensor-cut [135], a novel tensor-based graph-cut method based on the local neighboring Markov random field (MRF) model. The limitations of these methods lie in the user interaction and in the setting of different initialization parameters, required according to the input image and to the specific tubular structure to be segmented.

5.1.3 Deep learning-based methods for tubular structures segmentation

Moving on to deep learning-based works for tubular structures segmentation, we first want to focus on the method of Virzì *et al.* [132] which is not applied to renal ceCT images, but still dedicated to tubular structures segmentation. Authors propose a semi-automatic patch-based deep learning approach to segment pelvic vessels in 3D MR images of pediatric patients. To consider only relevant patches, the skeleton of the vascular tree is obtained combining user-selected landmarks and shape-appearance information, and then patches are extracted along this skeleton. This method also allows dealing with small volume to background ratio thanks to the possibility to create patches of smaller dimensions without losing information, and therefore at the same time to decrease training time and memory required by the CNN. This method works well for pelvic vessels whose branching is fairly simple and constant, although it always requires some user's interaction. In the case of renal structures, while the ureters tree is quite simple, this is not true for arteries and veins trees, as we can see in the example in Figure 5.3.

In order to overcome the problem given by the user's skeleton creation to work with only small and relevant patches, Dang *et al.* [28] propose the use of a PNet-based [138] patch classifier called Vessel-CAPTCHA. The difficulty of this method lies precisely in the training of the classifier, which requires a large number of patches to achieve good performance. For these reasons, the authors decided to work in 2D, losing the volumetric information, and to create a user-friendly annotation system on a large unlabeled training set.

In the following we focus on fully automatic methods.

Fully-automated algorithms A first fully-automated algorithm is Kid-Net [128]. It is a method to segment renal arteries, veins and collecting system via CNN on ceCT. The authors point out the higher importance to operate on these structures using 3D patches instead of 2D slices, since the latter do not have enough information. Moreover, instead of training for individual foreground classes independently, they train the network to detect the three

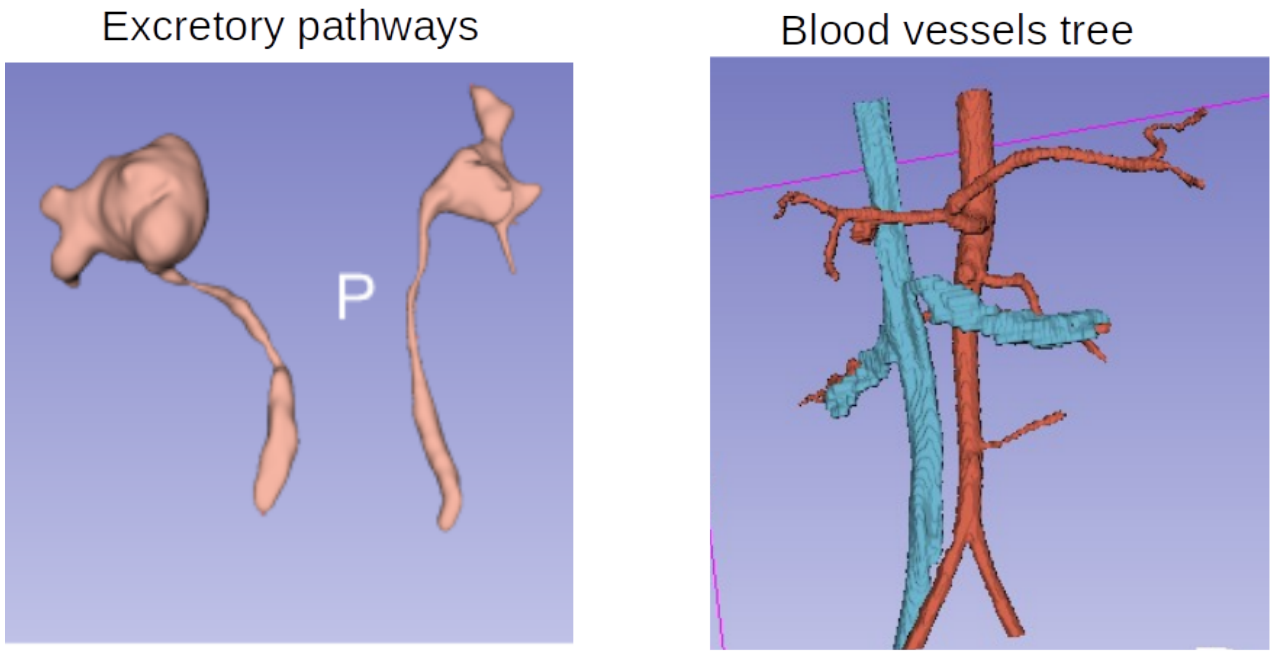


Figure 5.3: Ureters on the left in pink and renal blood vessels tree on the right with arteries in red and veins in blue.

foreground classes for the same reasons discussed in Section 3.1.2. Kid-Net is a 3D U-Net and its major contributions rely on the use of a deep supervision, a patch selection method that handles unbalanced data through the use of random sampling and a dynamic weighting based on volume to background ratio. They use cross-entropy as loss function. Results on 100 subjects of a private dataset of arterial phase ceCTs are above 85% of Dice score for arteries, degrading if we focus on the small vessels around the kidney, and around 60% for both veins and ureters because of greater difficulty due to the lower contrast intensity.

To tackle the difficulties on fine vessels segmentation and on having a small dataset, He *et al.* [51] propose DPA-HRA-DenseBiasNet. The DenseBiasNet takes standard 3D U-Net as the basic structure, but it compresses and transmits all feature maps in each layer to every forward layer. Moreover, the authors propose what they call “the deep prior anatomy (DPA) strategy”: an autoencoder (AE) is trained in an unsupervised manner with numerous unlabeled data in which noise is added and the AE is optimized to reconstruct the original image. Once the AE is trained, the features are embedded in the supervised model to guide it. Finally, they propose a hard region adaptation (HRA) loss function that samples the loss dynamically according to the segmentation quality of each pixel. The authors used a private dataset of 196 kidney cancer adults patients from which 392 patches are extracted and 156 of them were labeled by clinicians: 52 labeled patches were used for training the Dense 3D U-Net, the last 104 labeled patches for test set and the remaining 236 unlabeled images for training the AE. They achieve for the first time fine renal artery segmentation with Dice score equal to 88%, showing better results than Kid-Net and 3D U-Net trained and tested on the same database. It is important to underline that with the only use of the Dense 3D U-Net a Dice Score of 86% is already achieved, and the use of HRA and then HRA+DPA both add 1%. We would like to emphasize that no reason is shown why the use of the autoencoder for DPA leads to a priori anatomy modeling. The weaknesses of this method rely on the computing process that requires a lot of

memory and time, due to the dense connections of the network and the additional parameters of DPA strategy. Furthermore, this work has been tested only on arteries, experiments with veins and ureters still need to be done.

A faster method is presented in [85], in which the authors try to segment not only all tubular structures but also renal parenchyma and tumors with a single network. They use a Residual U-Net with a multi-scale weighted cross-entropy loss function that gives greater importance to foreground voxels, edges, and complicated and small structures. Tests on 100 ceCT images confirm the analysis by other authors with high Dice score for parenchyma (96%) and arteries (86%), acceptable for veins (80%), but a difficulty in segmenting the collecting system (62%) and tumor (29%).

Focusing now on methods not developed for the renal area but still focused on ceCT images, special attention should be paid to the well-known nnU-Net [61]. Among the various proposed contributions already presented in Section 3.1.2, similar to the Kid-Net [128] approach, the authors also propose the training of multiple “foreground” classes, the use of deep supervision and an oversampling technique. However this is different from Kid-Net’s one, we remind the reader in fact that nnU-Net ensures that one of the classes of interest is contained in at least 33.3% of patches in every mini-batch. The other 66.7% of patches is randomly selected from the entire training set. Another difference with Kid-Net lies in the final loss function, which is a combination of soft Dice loss and cross-entropy.

Other methods focused on ceCT images exploit distance map to improve performance. Distance maps are generated by computing the distance transform on the segmentation masks.

Some authors [102] propose to use it to leverage the differences in diameter of the vessels. In [134], authors use the inverse of the normalized distance map to weight the voxel of the reference segmentation in the Dice loss calculation. Their proposed loss function is called Radial Distance loss, *RDloss*, and successfully improves lung tree segmentation performances of both smaller vessels and voxels at the boundaries of thicker ones. A similar approach is proposed in [15] to segment tubular bones, such as femur and tibia, but the inverse distance map is applied using cross-entropy calculation.

This idea of exploiting the geometry of the vessels is also used by Wang *et al.* [139], who propose to include the tubular geometry information directly in the training of a CNN. The distance map is created for every manual segmentation, and used as reference for a loss function with the second output channel of the network (where the first output channel is the segmentation mask). During inference, the segmentation mask is refined by leveraging the shape prior reconstructed from the distance map (named Geometry-Aware Refinement, GAR). One important proposition is that the distance map D is quantized on values from 0 to K (calculated as the maximum possible distance) and the cross-entropy is calculated between the discrete distance map D and the probability that a voxel belongs the k -th class (softmax of the second output with $K + 1$ channels). In this way, the authors [139] formulate the distance prediction problem as a classification problem. The authors explain that the use of distance map without quantization can make the training difficult because outliers can cause large errors and lead to unstable predictions (resulting in difficulty for the network to converge). Experiments showed better performances compared to the previous state-of-the-art methods on aorta, cava vein and hepatic vessels segmentation in ceCT scans, automatically providing also the distance map for a geometrical measurement of the tubular structures, at the expense of high computational time and memory. The limitation of this method also lies in the possibility of segmenting one structure per network.

Recently some authors are beginning to combine CNN with Graph Neural Network (GNN) with satisfactory results for tasks such as lung vessel-tree semantic segmentation in ceCT images [41, 129] and head and neck artery semantic segmentation in angiographic ceCT images [148]. The CNN gives a rough estimate of branch endpoint landmarks and binary branch segmentation locations, and then the GNN refines these rough estimations to produce the final semantic segmentation. However, their utility lies in being able to do semantic segmentation from binary labels, i.e. segmenting lung vessel-tree branches to fine categories of n segments while training the network with the entire lung vessel-tree, which is not our purpose. Moreover, CNN-GNN networks require a very high computational cost as well as a high amount of input images to converge.

5.1.4 Assessment considerations

According to what has been analyzed, to date there is still no method capable of segmenting all renal tubular structures in arteriovenous ceCT images with high performance, in addition to the fact that none of those presented used pediatric ceCT images, which present further difficulties. However, various techniques presented in this section appear to be the basis for good performances, such as the use of a single network to detect multiple adjacent structures, the deep supervision, and the oversampling and patch selection method. These techniques allow to overcome certain limitations related to the segmentation of tubular structures, including the small available labeled dataset, the intra-scale changes, and the small volume to background ratio. The main limitation still seems to be related to the loss functions used. In fact, voxel-wise functions do not seem sufficient for tubular structures segmentation, and the combined use of distance map does not seem to be effective, presenting instead a redundancy of information for the network. Taking inspiration from non-machine learning methods, and in particular from vesselness filters, we propose instead to combine the voxel-wise losses with a new loss based on the eigenvalues of the Hessian matrix and on the vesselness function itself. Our idea is to exploit both the *a priori* structural morphological knowledge of tubular structures and the information about their neighborhood, which must be continuous in the main structure direction.

5.2 Methods

In this section, we first present the approaches for selecting and implementing the methods we will compare. Then we propose a new loss function based on the vesselness. We show that adding this loss function to a merge of the best method found and a proposed oversampling method, allows further improving both qualitative and quantitative performances.

5.2.1 Comparison of state-of-the-art methods of renal tubular structures segmentation in ceCT images

We focus on deep learning-based methods due to the difficulties of rule-based methods related to heterogeneity of image intensity in arteriovenous ceCTs, as shown in Section 5.1.2.

Methods selection

We decided to compare some of the methods presented in Section 5.1.3, summarized in Table 5.1. Our selection criteria are:

- (i) fully-automated algorithm,
- (ii) distinctive peculiar techniques,
- (iii) code available or easily reproducible.

The first criterion was chosen as a result of the problems set out in Section 5.1.3, related to the works of Virzì *et al.* [132] and Dang *et al.* [28]. For the second criterion, we have identified three distinctive techniques: the use of the distance transform as a second loss term, the dense connection method useful for segmenting fine structures, and the deep supervision for the loss calculation. Among the methods based on the previous peculiar techniques, almost none have the code available online but some are easily reproducible, given the many details provided by the authors. We selected the method proposed by Wang *et al.* [139] for distance transform (Method 1), the method of He *et al.* [51] for the dense connections (Method 3), and Kid-Net [128] for the use of deep supervision (Method 4). A variant of the Deep Distance Transform method presented by Ma *et al.* [90], whose code is available online, has also been analyzed (Method 2). Lastly we selected the nnU-Net [61], as it is a high performing method in general for the segmentation of medical images with its deep supervision method (such as Kid-Net) optimizing CE combined with Dice score (Method 5). Moreover, nnU-Net code is available online for the purpose of benchmarking, in contrast to the similar Kid-Net algorithm.

Table 5.1: 3D fully-automated supervised methods selected for comparison

N°	Method	Backbone	Outputs	Loss functions
M1	Deep Distance Transform [139]	U-Net (depth 5)	1: Segmentation 2: DistanceMap One-Hot Encoder	1: SoftmaxCE 2: SoftmaxCE
M2	Deep Distance Transform [90, 139]	U-Net (depth 5)	1: Segmentation 2: DistanceMap	1: SoftmaxCE 2: Mean of L1
M3	DenseBiasedU-Net [51]	DenseU-Net (depth 4)	1: Segmentation	1: SoftmaxCE+Dice
M4	Kid-Net [128]	U-Net (depth 5)	1: Segmentation	1: SoftmaxCE w/ deep supervision
M5	nnU-Net [61]	U-Net (depth 5)	1: Segmentation	1: SoftmaxCE+Dice w/ deep supervision

Methods implementation

We implemented the selected methods (summarized in Table 5.1) as follow:

- M1. We implemented this code starting from the available code of Method 2 which can be a variant of this. To be able to apply the idea of Deep Distance Transform method [139] to formulate the distance prediction problem as a classification problem (as explained in Section 5.1.3) to two structures (i.e. arteries and vein), we propose that the discrete distance map of veins is added to the discrete distance map of arteries. We add K_A (maximum possible distance of the distance map of arteries) to all values of the distance map of veins where different from 0 (the distance map is 0 where the reference segmentation is 0 so this idea cannot produce errors). In this way the final distance map D_{map}

presents values from 0 to $K_A + K_V$, where K_V is the maximum possible distance of the distance map of veins. By doing so, after quantization, the final distance map D_{map} as well as the second output will have $K_A + K_V + 1$ channels. We coded also the Geometry-Aware Refinement (GAR) presented in [139], in order to refine the segmentation output with the quantized distance map output.

- M2. This is a variant of the Deep Distance Transform method [139] that was implemented by Ma *et al.* [90]: the second channel, i.e. the distance map, is not quantized. Authors [90] used an L1 norm as loss between the reference distance map D_{map} and the second output of the network (Conv3D $1 \times 1 \times 1$ with no activation function). This implementation makes it easier to use the distance map for multiple structures but does not respect the idea of the original paper of Wang *et al.* [139] (Method 1). The code is available online.
- M3. We implemented the DenseBiasedU-Net [51] as in the original paper from scratch. Each dense biased connection compresses via a convolutional layer the feature maps in each layer of the U-Net at only 4 feature maps. Then, these are transmitted and concatenated to every forward layer. The reduction is done in order to reduce feature redundancy while keeping the integrity of information flow and gradient flow, allowing also to fuse multi-scale features. In order to fit the network in a 16 GB GPU, the 3D U-Net has depth of one layer less than the other methods. The HRA and DPA techniques were not implemented due to their low contribution in improving performance (see Section 5.1.3) and the limited database at our disposal (see Necker PRAC database at Chapter 2).
- M4. We implemented this code starting from the available code of Method 5, using only CE in the loss function and with the different oversampling method presented in Section 5.1.3.
- M5. The original implementation is available online for nnU-Net [61]. Here we used our high fidelity implementation from scratch detailed in Section 3.2.

5.2.2 Proposed tubular structures loss function

Motivation

The use of Dice score and cross-entropy revealed to be not enough to evaluate the segmentation performance on fine tubular structures. Both are very sensitive to small structures: changing a few voxels can change the score significantly, in particular the Dice score. This can also affect the training process with patches (example in Figure 5.1): we can have high gradient even when the number of wrong pixels is small, and additionally we can have fluctuations due to very different batches. Furthermore, when large and small blood vessels are present in the same patch (e.g. aorta or cava vein and very tiny renal vessels), we will have a good Dice score but the algorithm only segments the aorta or cava vein, while it completely misses the small renal vessels which represent a lower percentage of the foreground voxels in the patch. Eventually, due to the heterogeneity of the vessels, there is a strong uncertainty about some vessels segmentation which may result in prediction of vessels with interruptions.

However, voxel-wise information is necessary to perform voxel classification, such as segmentation. Moreover, for medical image segmentation, cross-entropy proved to be not enough to reach high performance [61], due to the extreme scarcity of foreground voxels in a patch,

that will force the network to have a strong bias to the background [155]. The use of the Dice score in the loss function tackles this problem, carrying with it the limitations outlined earlier.

In order to introduce information that is exempt from the number of pixels of the reference segmentation and that also is not voxel-wise but takes the neighborhood into account, we propose to leverage the use of the eigenvalues of the Hessian matrix and the vesselness function, described in Section 5.1.1, as a loss function for the segmentation masks. In fact, on the one hand, the eigenvalues allow us to verify that the structural morphology of the predicted segmentation is similar to that of target structure. On the other hand, the vesselness allows us to enhance the segmentation of elongated structures without interruptions. Finally, the different sizes of vessels in a patch can be taken into account by using such loss functions in a multi-scale manner via deep supervision.

Due to the different images, namely the segmentation masks, to which these functions are applied, a new formulation of the steps for calculating eigenvalues and the resulting vesselness score is presented in the next section.

Formulation

As presented in Sections 5.1.1 and 5.1.2, the application of vesselness functions on abdominal ceCT images with arteriovenous phase leads to unsatisfactory results. Taking inspiration from the use of vesselness for registration presented in Section 5.1.1, here we propose for the first time to translate the use of vesselness as loss cost function for segmentation purpose. Therefore, our proposed vesselness function will not be applied on the abdominal ceCT input image but on its segmentation mask and on the predicted one. These images in fact exhibit the sought-after characteristics for a satisfactory vesselness application, similar to those of the lung ROI (bright tubular structures in black background) used in [16] and [136].

One might argue that using such a function on the product between input and segmentation mask would be more appropriate; however, the heterogeneity of ceCT images, particularly of the pediatric ones, makes it complicated to find vesselness parameters that are appropriate for all the structures present in the patch.

Furthermore, since the probability map at the output of the network is not a binary object, an approach directly adapted to binary objects (e.g. moments comparison) would not be appropriate. In fact, our idea stems from the analysis of probability maps at the network output of a classic 3D U-Net: in these, tubular structures were found, but because of the problems exposed in the previous paragraph, a low probability to the correct class was assigned to contour voxels or to the ones belonging to finer portions of the vessel. We noticed that the use of Frangi's vesselness increases the probability assigned to the voxels of each class that most respects the vesselness. Figure 5.4 shows this more clearly.

However, applying it as post-processing on probability maps may greatly increase false positives. For this reason, we incorporated this idea directly into the training of the neural network as a loss function. In order to do this, we need to transform the reference segmentation, which is instead a binary object, as applying the Hessian matrix on this will result on a non-zero gradient only on the edges. The proposed pipeline for our Hessian-based vesselness is as follows:

1. **Gaussian filtering.** We smooth both the binary segmentation of the reference and the probability map of the prediction by applying a convolution with a Gaussian kernel strong enough to have zero-gradient only in the principal dimension. Nevertheless, a

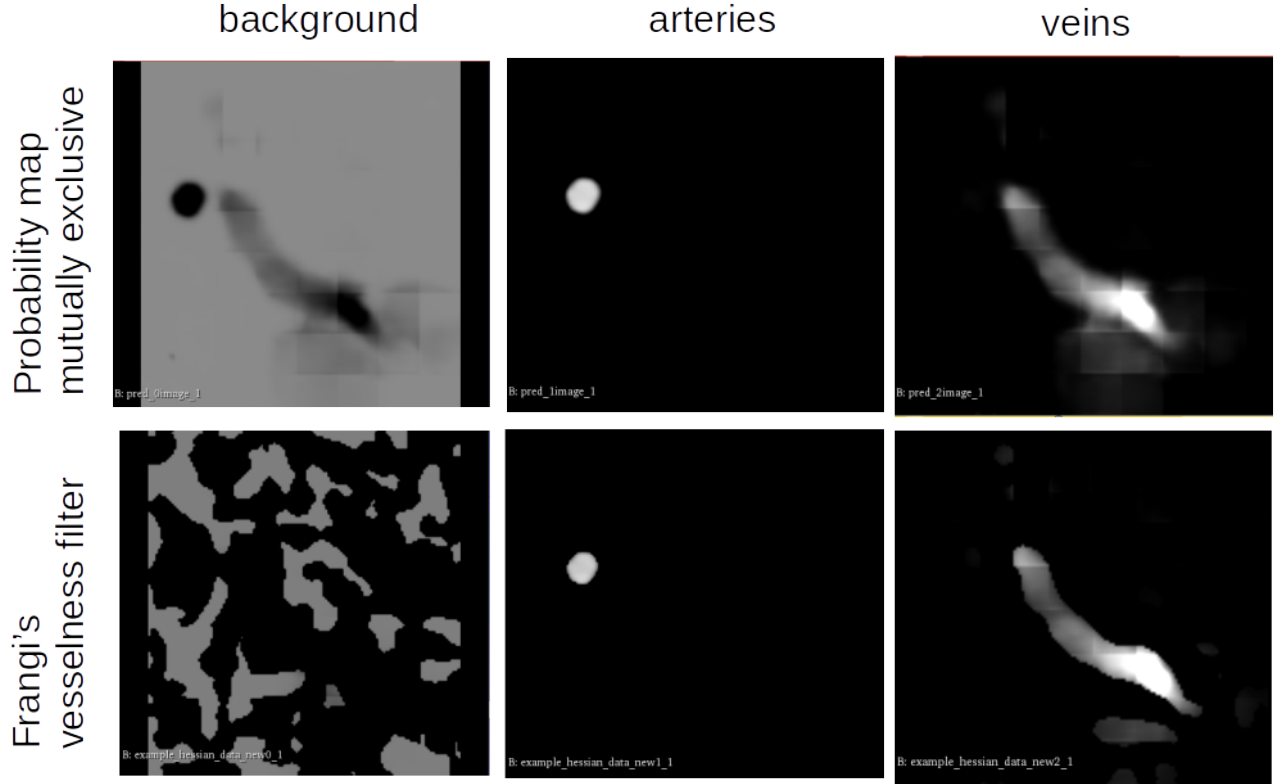


Figure 5.4: First row: map of probability by class as output of the network after using a softmax function. Second row: Application of Frangi's vesselness filter. Color code: gray scale from black as 0% of probability of a voxel to belong to that class and white as 100% of probability.

high standard deviation σ might make the small blood vessels completely disappear. As a compromise, taking inspiration from [38], we apply 5 different Gaussian kernels with $\sigma_{i \in [1,5]}$ ranging from 1 to σ_{max} with a step of $\frac{\sigma_{max}}{5}$. The value of σ_{max} is found empirically, in relation to the size of both the selected patch and the structures to be segmented. The final smoothed predicted P (or reference R) segmentation is computed as: $P_{g\sigma} = \sum_{i=1}^5 (g_{\sigma_i} * P)$. This way we ensure that we only have zero gradient along the main direction.

2. **Hessian matrix calculation.** We calculate the Hessian matrix as in Equation 5.1 for every voxel of the filtered segmentation $P_{g\sigma}$, as $H(\sum_{i=1}^5 (g_{\sigma_i} * P))$. It is important to emphasize that thanks to the convolution with the Gaussian kernel presented in the previous step, we ensure that the second partial derivatives for each voxel of the segmentation masks are all continuous and that each Hessian matrix is a symmetric matrix by Schwarz' theorem. This is fundamental because the computation of eigenvalues is differentiable only for real symmetric matrices [8, 91, 143]. Further details on the differentiability are provided in Appendix E.
3. **Ordering of eigenvalues.** Due to the fact that in the case of predictions which include initially no structure or structures with different shapes and directions, using directly the vesselness function could result in a training slowdown or even in worse segmentation performance. This is because such vesselness functions require to sort the eigenvalues by magnitude, that in our case could end in prediction P and reference R having for

the same voxel v a very similar vesselness score for structures with different preferential directions. In order to overcome this issue, we order the eigenvalues of predicted and reference voxels (resp. P_v and R_v) via their associated normalized eigenvectors W . In particular we match the eigenvectors with the smallest angle between them, namely the minimum rotation required to overlap them, finding $\arg \min_{\mathcal{P}} \sum_{i=1}^3 \|(W_{P_v} \mathcal{P})_i - (W_{R_v})_i\|_2$, where i is the index of the column representing the associated normalized eigenvector and \mathcal{P} are all the possible permutation matrices. Figure 5.5 illustrates this idea.

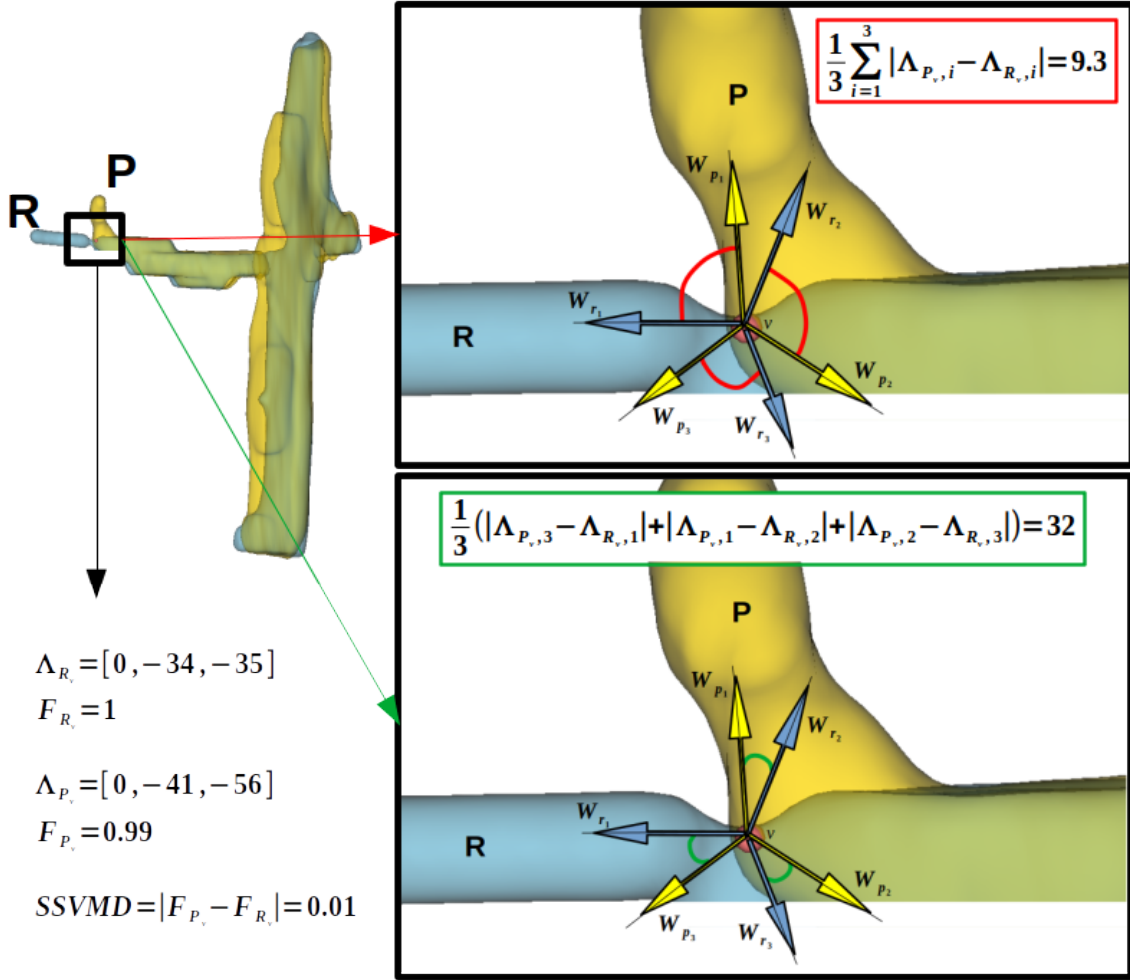


Figure 5.5: Ordering of eigenvalue vectors Λ of the same voxel v in order to allow for a fair comparison between them. Using SSVMD to compare vesselness score F results in a very low value even if prediction P (yellow) and reference R (light blue) have different main directions. Moreover, a comparison of eigenvalues ordered by magnitude does not reflect the real dissimilarity between the two segmentations (top box on the right). In order to overcome this issue, we order the eigenvalues of predicted and reference voxels via their associated normalized eigenvectors W , matching the W with the smallest angle between them (bottom box on the right). This matching is important in this case, where a voxel-wise loss function would fail to correctly assess this error due to the fineness of the portion of the vessel.

4. **Multi-scale supervision.** In order to inject as much information as possible to the network, we do the same for the subsequent three output levels of resolution using the above mentioned deep supervision technique in Equation 3.3. However, given the lower spatial definition of these outputs, the number of σ values used for the Gaussian kernel

is set to $5 - q$, i.e. $\sigma_{i \in [1, 5-q]}$, where q is the resolution level as in Equation 3.4 (0 as first level).

As motivated in the previous paragraph, our *vesselness* loss function, named Tubular structures Loss - *TsLoss*, is composed of two parts: a first loss function to check the morphology of the structures, named morphological similarity loss function and denoted by *MsLoss*, by comparing the eigenvalues ordered by the eigenvectors matrix as presented above; a second loss function to force prediction of elongated structures as in Frangi's vesselness function, and thus named Frangi vesselness loss function *FvLoss*.

The **morphological similarity loss function** is defined for a single image and a single target structure at level of resolution q as:

$$MsLoss_q(P_q, R) = \frac{1}{3M_{\hat{R}}} \sum_{m=1}^{M_{\hat{R}}} \sum_{o=1}^3 (\Lambda_o(H(\sum_{i=1}^{5-q} (g_{\sigma_i} * p_m))) - \Lambda_o(H(\sum_{i=1}^{5-q} (g_{\sigma_i} * r_m))))^2 \quad (5.5)$$

where g_{σ_i} are the $5 - q$ different Gaussian filters applied to the segmentation masks, with standard deviations σ_i (as explained before in step 1), H is the Hessian matrix (step 2), Λ is the array containing the three eigenvalues of H ordered by the associated normalized eigenvectors (via the smallest angle as explained before in step 3 and Figure 5.5), p_m is the probability of a voxel m of the predicted segmentation P (i.e. the output probability map) at resolution q and r_m is the corresponding target sample of the reference segmentation R , while $M_{\hat{R}}$ is the number of voxels of the dilation of R with a square structuring element of size $3 \times 3 \times 3$ (calculating eigenvalues over the entire image is expensive in terms of computational time, and the use of dilation revealed to be sufficient for our purpose thanks also to the combined use of voxel-wise loss functions). Moreover, this loss function allows us to take in consideration also flattened and deformed vessels (due to the presence of the tumor), in which instead the direct use of Frangi's vesselness may not be useful due to vesselness scores that can be very close to 0 and thus too similar to the score of *non-found* vessels. In the global loss function, the *MsLoss* term is weighted by a factor w_s .

The **Frangi's vesselness loss function** is designed in a non-supervised way for a single image and a single target structure at level of resolution q as:

$$FvLoss_q(P_q, R) = \frac{1}{M_{\hat{R}}} \sum_{m=1}^{M_{\hat{R}}} (1 - \frac{F(H(\sum_{i=1}^{5-q} (g_{\sigma_i} * p_m)))}{F_{max}}) \quad (5.6)$$

where $M_{\hat{R}}$ is the number of foreground voxels of R , F is the Frangi's vesselness presented in Equation 5.3, F_{max} is the maximum among the $M_{\hat{R}}$ Frangi's vesselness values. This loss function forces voxels corresponding to the target structure to have a high vesselness value, avoiding the vanishing gradient problem. The use of the former *MsLoss* function allows forcing the correct direction of the predicted vessels and thus enables the possibility of using Frangi's method without favoring incorrect predictions.

The complete **tubular structures loss function** is:

$$TsLoss = \frac{1}{N} \sum_{n=1}^N \sum_{q=0}^Q w_q \cdot \frac{1}{C} \sum_{c=1}^C (w_{ms} MsLoss_q(P_{n,q,c}, R_{n,c}) + FvLoss_q(P_{n,q,c}, R_{n,c})) \quad (5.7)$$

where N is the batch size, C the number of structures to be segmented (not counting the background), $P_{n,q,c}$ is the prediction P_q for the class c of the image n of the batch, $R_{n,c}$ is

the corresponding reference segmentation, and Q the number of output resolution levels taken into account (equal to 4 in our experiments, as explained before in step 4) with w_q as in Equation 3.4. This loss function is added to the voxel-wise functions, such as cross-entropy and soft Dice loss terms.

Figure 5.6 shows the complete pipeline for the proposed tubular structures loss function.

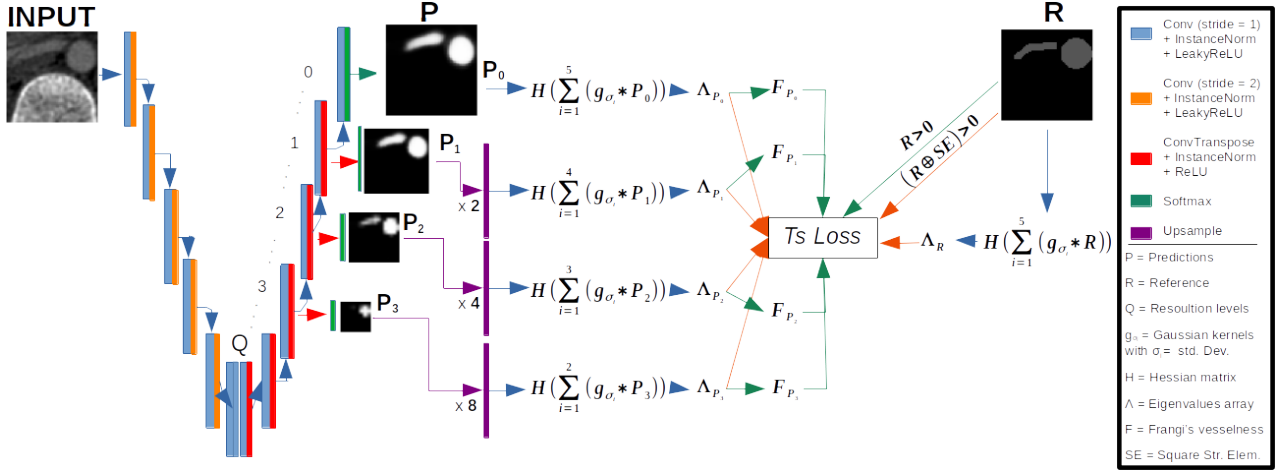


Figure 5.6: Tubular structures loss function ($TsLoss$) pipeline. This function is composed of two loss terms, one that uses the eigenvalues Λ of the Hessian matrix calculated for every voxel of the filtered segmentation masks, and the other that uses the F Frangi's vesselness function [38] calculated from these eigenvalues. The $TsLoss$ term is applied in deep supervision for the first 4 resolution levels, together with voxel-wise loss functions. See text for details.

5.3 Materials and Experiments

We present two sets of experiments. The first one has the aim to segment arteries and veins with the same network for the reason presented in the previous section. The second set of tests has the goal to train a network to only segment the ureters. The three structures are not segmented with the same network because they are labeled on different ceCT acquisition modalities, as detailed in Chapter 2.

5.3.1 Database

We worked on the Necker PRAC database presented in Chapter 2.

For ureters segmentation we trained and tested on the 16 ceCT scanners with excretory/delayed phase. In particular we trained the networks using 10 patients, keeping 1 for validation. For inference we used the data from 5 patients: 2 with mono-phasic injection and the 3 with bi-phasic injection.

For the segmentation of arteries and veins we trained on the 63 ceCT images with vascular/early phase of mono-phasic injection (see Figure 1.4). In particular, we used 46 ceCT for training and 5 for validation. For testing we used the images of 15 patients: 12 with mono-phasic injection and 3 with bi-phasic injection.

All images are pre-processed as for the nnU-Net [61] (and as done in the previous chapters, see Section 3.4.1) and are pre-cropped in the abdominal ROI (using the automatic method

and on-line method presented in Chapter 4.1). Eventually, images are divided into 3D patches of size $96 \times 160 \times 160$.

5.3.2 Training implementation details

The number of training epochs is fixed at 1000 with 500 patches seen at each epoch, randomly chosen from the training set. The number of iterations at each epoch depends on the mini-batch size, specifically set to 4 in our experiments. All trainings and tests were run using the same specifications as presented in Section 3.4.1. Patches are randomly extracted from the abdominal ROI for training and validation images. During inference we operated as in nnU-Net [61] using the sliding window on the abdominal ROI with overlapping of half of the size of the patch and the Gaussian importance weighting.

Due to the strong imbalance between patches with only background and patches with structures, an oversampling technique for selecting at least 50% of patches with a minimal number of voxels per structure, *MinPix*, was adopted for all methods in Table 5.1 for which no oversampling was done (M1 - M3). This choice was made because early results without such a technique had very poor performance for veins and ureters. Moreover, for networks that segment arteries and veins, an additional oversampling on patches with structures is done due to the lower presence of the latter, ensuring that the previous 50% are equally distributed. A sufficient *MinPix* was empirically found as 1000 voxels for all the structures. For Kid-Net (M4) and nnU-Net (M5) we used the oversampling technique of these methods, previously presented.

The on-the-fly spatial and iconographic data augmentation presented and discussed in Section 3.1.2 is applied at each iteration (details in Appendix B). In this case the augmentation is applied to both the entire input images and the target structures alone (only the iconographic changes), in order to better manage heterogeneity of image intensity. As discussed in Chapter 3, this step is critically important when working on 3D patches (as opposed to work on 2D slices, as demonstrated). The use of synthetic CT images to tackle heterogeneity problems is not used in this chapter in order to evaluate the techniques independently. The results obtained with the method proposed in this chapter combined with the use of both real ceCT and synthetic CT are shown in Chapter 6.

Stochastic Gradient Descent (SGD) with a Nesterov momentum of 0.99 is used as optimizer for all tests. The initial learning rate lr of 0.01 is reduced following the poly learning rate policy [18], decaying at each epoch e by multiplication of the initial lr by a factor of $(1 - \frac{e}{E})^{0.9}$ where E is the total number of epochs. In our proposed method, we also assessed the performance of Adam [75] and Adagrad [34] optimizers with initial lr of 10^{-3} , taking inspiration from the work in [78]. This study is described in Appendix E.

We empirically found the best value of weight w_{ms} of *MsLoss* in Equation 5.7 as 0.05 for arteries and veins, and 0.01 for ureters, and the weights of F (Equation 5.3) for *FvLoss* (Equation 5.6) as $\alpha = 0.1$, $\beta = 0.1$ and $\gamma = 2$. We also found the best σ_{max} to ensure zero gradient only in the main direction equal to 25. The search of these parameters is presented in Appendix E.

In addition to *TsLoss*, for the reasons stated in Section 5.2.2, cross-entropy (CE) and Soft Dice score (to which we will refer simply as Dice in tables and figures) were used in the loss function for trainings.

5.3.3 Evaluation measures

For the quantitative evaluation of the segmentation results, we compute three different categories of measures. The first two are the same as in the previous chapters and detailed in Appendix A: Dice score, precision and recall as spatial overlap based measures, and the 95th percentile of Hausdorff distance as spatial distance. However, these measures carry with them the limitations presented in the previous sections, and for this reason, alone they are not sufficient for a proper evaluation of the segmentation results. To overcome these limitations, we decided to use also our proposed morphological similarity loss (*MsLoss* in Equation 5.5) for the motivations discussed in Section 5.2.2. In the result tables we refer to this measure as $\Delta\Lambda$.

Moreover, we perform a further analysis for arteries and veins, which we refer to as Recall analysis. In this study, arteries and veins of the reference segmentation are semantically segmented into substructures that differ in diameters and directions. Arteries are divided into aorta, renal arteries and celiac artery, and veins into cava vein and renal veins. Aorta and cava vein are larger and follow approximately a constant direction, renal arteries and veins are very tiny vessels with irregular directions, while the celiac artery has a medium diameter between the previous structures and has a T-shape that branches perpendicular from the aorta on the coronal plane. An example of this sub-division is shown in Figure 5.7. The more a vessel is fine and irregular, the more the difficulty in manual segmentation increases. For this reason, the Recall measure between the prediction and each of these parts of the vessel tree is calculated. Indeed, we believe that having fewer false negatives is really important to speed up the 3D anatomical modeling process, since manual segmentation of missing parts takes longer than false positive removal. This is due to the proximity of arteries and veins and the fineness of the tubular structures being segmented.

5.4 Results and Discussion

In this section we show and discuss the results obtained with the methods of Table 5.1 and with our proposed method. A further study on different implementations of the proposed vesselness loss functions in both the best method found in the comparison and the proposed method is detailed in the Appendix E.

5.4.1 Arteries and veins segmentation

Table 5.2 shows the quantitative results for the comparison of state-of-the-art methods for arteries and veins using the evaluation measures presented in the previous section. The Deep Distance Transform method [139] (M1) outperforms the other state-of-the-art methods in all the measures (precision and recall need to be read together). Nevertheless, once the GAR post-processing is removed (which takes a long time in inference), the performances drop. The DenseBiasedU-Net [51] (M3) reduces false positives, as we can see from the values of Precision and 95th of Hausdorff Distance (95HD). The problem with this technique lies in the lower depth of U-Net, due to the large amount of memory required by dense connections, which limits the network’s information extraction. For what concerns Kid-Net method [128] (M4), we cannot say if the worse results are due to the oversampling technique or the only use of cross-entropy as loss function. Given the high presence of false negatives (low recall

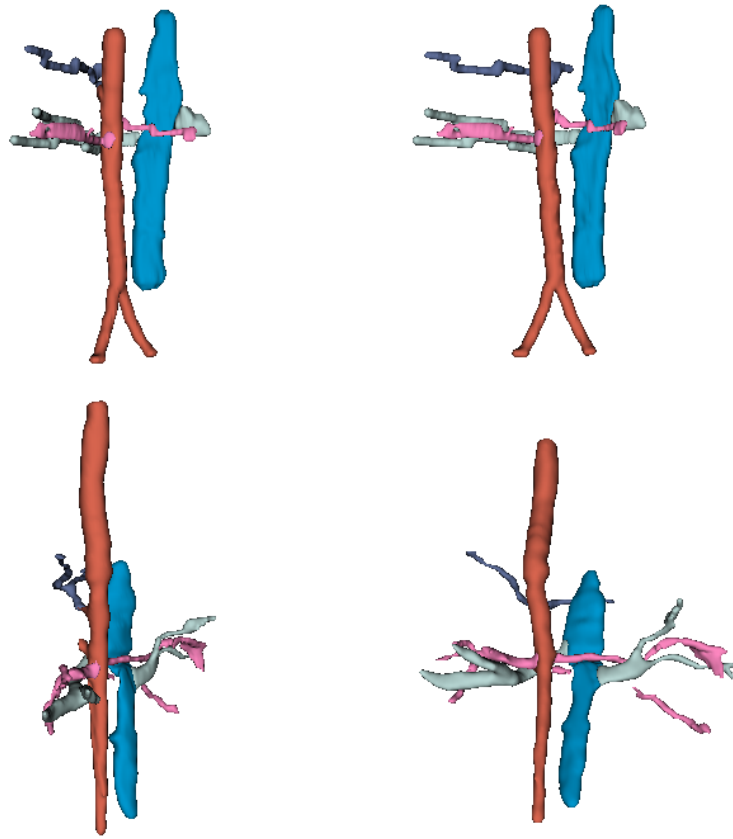


Figure 5.7: Examples of sub-division of the reference segmentation for the Recall analysis. Arteries are divided into aorta (in color red), renal arteries (fuchsia) and celiac artery (purple), while veins in cava vein (blue) and renal veins (light blue). Left: lateral view. Right: front view.

in Table 5.2) we think it may be due more to the latter. Overall, nnU-Net [61] (M5) can be identified as the best technique, where deep supervision with Dice and CE leads to good results without any heavy post-processing. For this reason and because of the ease in building on this method, we decided to apply our oversampling method within nnU-Net and then add our proposed loss functions. The proposed oversampling method is effective and comparable to the one presented in nnU-Net, nevertheless it seems to better balance the number of voxels examined for each class in the case of multiple structures, reducing the difference in performance among classes. The use of $MsLoss$ greatly improves segmentation results, highly reducing false negatives and both spatial distance and morphological similarity between prediction and reference. The addition of $FvLoss$ to build the final $TsLoss$ decreases even more the number of false negatives, at the expense of increasing false positives, but with significant morphological similarity improvement.

This is best seen from the qualitative results in Figure 5.11 (worst, average and best results for each method) and from the Recall analysis in Figure 5.8. In the latter, we can infer that the use of the proposed loss functions allows the network to identify better vessels of smaller diameter and with different directions (and thus morphology) from each other. Moreover, better results for both vein sub-structures underline that such vesselness loss functions may overtake heterogeneity problems.

We underline that the use of Dice and cross-entropy (CE) alone does not consider the

Table 5.2: Results on 15 patients using bbox patches $96 \times 160 \times 160$ obtained with the methods of Table 5.1 and with our proposed method for arteries (A) and veins (V) segmentation. *without GAR. **without HRA and DPA. Mean and standard deviation of the results are given. Other quantitative results on arteries and veins segmentation using patches of smaller size and different loss function implementations are shown in Appendix E.

Method	Oversampling	S	Dice Score [%] (\uparrow)	Precision [%] (\uparrow)	Recall [%] (\uparrow)	95HD [mm] (\downarrow)	$\Delta\Lambda$ (\downarrow)
M1 - Deep Distance Transform* [139]	Proposed <i>MinPix</i>	A	69.99 (3.44)	85.08 (10.63)	60.42 (5.28)	16.65 (13.29)	21.62 (1.99)
		V	37.29 (23.06)	82.59 (19.62)	26.77 (19.69)	27.28 (26.41)	22.51 (3.11)
M1 - Deep Distance Transform [139]	Proposed <i>MinPix</i>	A	71.91 (3.85)	81.39 (10.44)	65.47 (5.42)	16.40 (13.45)	20.19 (1.95)
		V	41.98 (23.76)	78.55 (18.95)	32.32 (22.41)	27.57 (29.59)	21.78 (2.95)
M2 - Deep Distance Transform [90, 139]	Proposed <i>MinPix</i>	A	63.73 (5.63)	78.52 (13.59)	54.56 (4.75)	22.75 (13.95)	22.44 (1.79)
		V	32.26 (22.60)	82.26 (26.92)	22.21 (16.81)	36.29 (20.49)	22.28 (3.25)
M3 - DenseBiasedU-Net** [51]	Proposed <i>MinPix</i>	A	65.76 (4.15)	86.95 (11.79)	53.71 (4.64)	18.26 (15.17)	23.42 (1.95)
		V	34.89 (22.20)	85.86 (10.37)	24.66 (18.93)	19.62 (10.10)	23.07 (3.09)
M4 - Kid-Net [128]	Random and dynamic weighting	A	65.70 (3.23)	88.93 (9.88)	52.82 (5.29)	19.48 (7.88)	23.19 (1.92)
M5 - nnU-Net [61]	Foreground in 33.3% of mini-batch	V	28.36 (23.56)	87.01 (16.84)	19.78 (18.87)	28.52 (21.23)	22.02 (3.89)
		A	68.05 (5.26)	84.43 (11.05)	57.85 (6.32)	15.55 (9.13)	22.12 (1.99)
Proposed 1: Deep Sup. Dice + CE	Proposed <i>MinPix</i>	V	39.78 (16.80)	79.36 (12.87)	28.15 (14.27)	25.12 (28.97)	23.18 (2.60)
		A	63.45 (5.67)	71.73 (9.99)	57.87 (7.31)	17.46 (9.65)	21.15 (1.93)
Proposed 2: Deep Sup. Dice + CE + MsLoss	Proposed <i>MinPix</i>	V	42.64 (20.12)	76.67 (13.17)	31.84 (17.12)	23.55 (17.00)	21.38 (3.27)
		A	75.88 (3.03)	87.92 (4.64)	67.09 (6.15)	9.79 (4.58)	19.67 (2.39)
Proposed 3: Deep Sup. Dice + CE + TsLoss	Proposed <i>MinPix</i>	V	60.33 (25.63)	81.76 (9.33)	53.26 (27.28)	18.65 (21.29)	19.47 (3.65)
		A	76.77 (3.93)	80.41 (10.17)	75.04 (7.66)	10.02 (5.80)	17.73 (2.49)
Proposed 3: Deep Sup. Dice + CE + TsLoss	Proposed <i>MinPix</i>	V	58.35 (26.79)	75.83 (13.01)	54.09 (28.87)	18.84 (21.31)	19.37 (3.29)
		A					

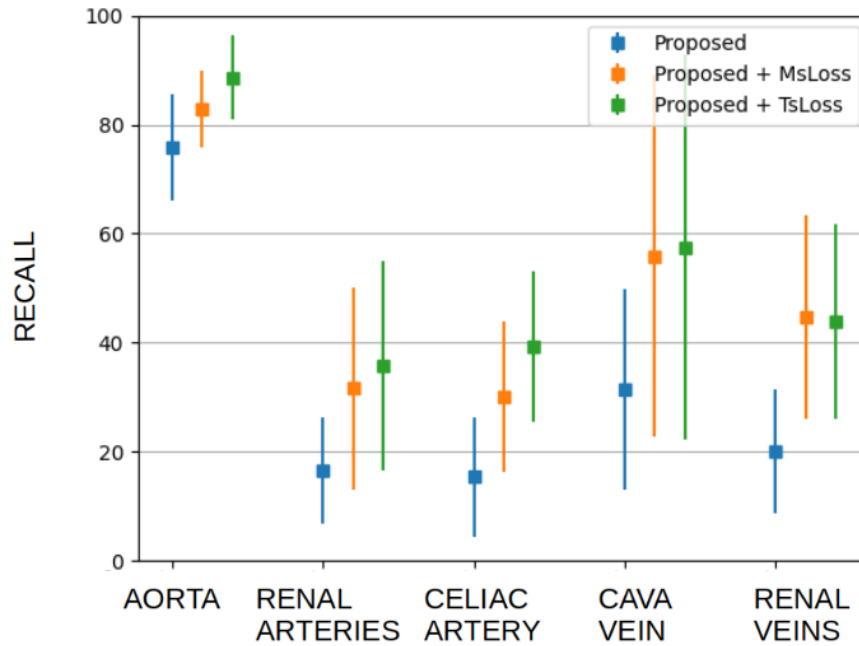


Figure 5.8: Recall analysis of the last three rows of experiments in Table 5.2 for the different structures. Arteries are divided into aorta, renal arteries and celiac artery, while veins in cava vein and renal veins.

minimization of morphological differences, as can be seen from the graphs in Figures 5.9 and 5.10. In these experiments we selected 100 patches with high presence of the target structures and training with different combinations of loss functions for 500 epochs. The *TsLoss* and *MsLoss* values are plotted for each training, including one in which it is not used as a loss function during training (in orange). In Figure 5.9, the proposed loss function is considered in its entirety as *TsLoss* and we can see that it is not minimized during Dice+CE

training. In Figure 5.10, the attention is on the morphological loss term ($MsLoss$), and we can notice that the use of the specific vessellness loss term ($FvLoss$) helps in better optimizing the $MsLoss$ (red and green curve).

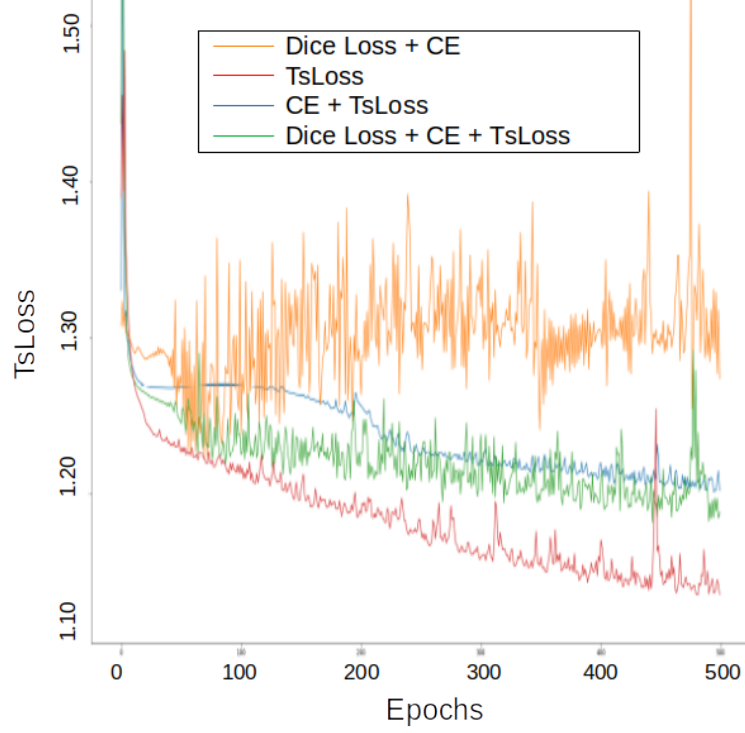


Figure 5.9: How $TsLoss$ behaves during training for different combinations of loss functions.

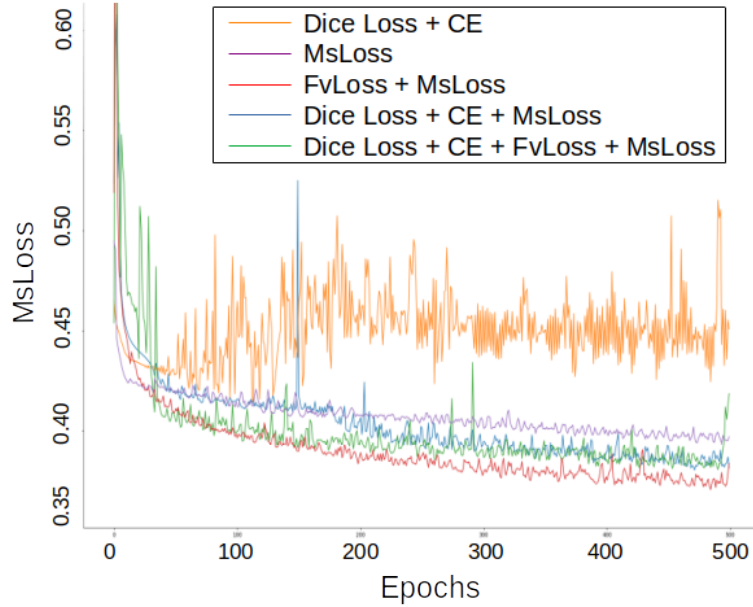


Figure 5.10: How $MsLoss$ behaves during training for different combinations of loss functions.

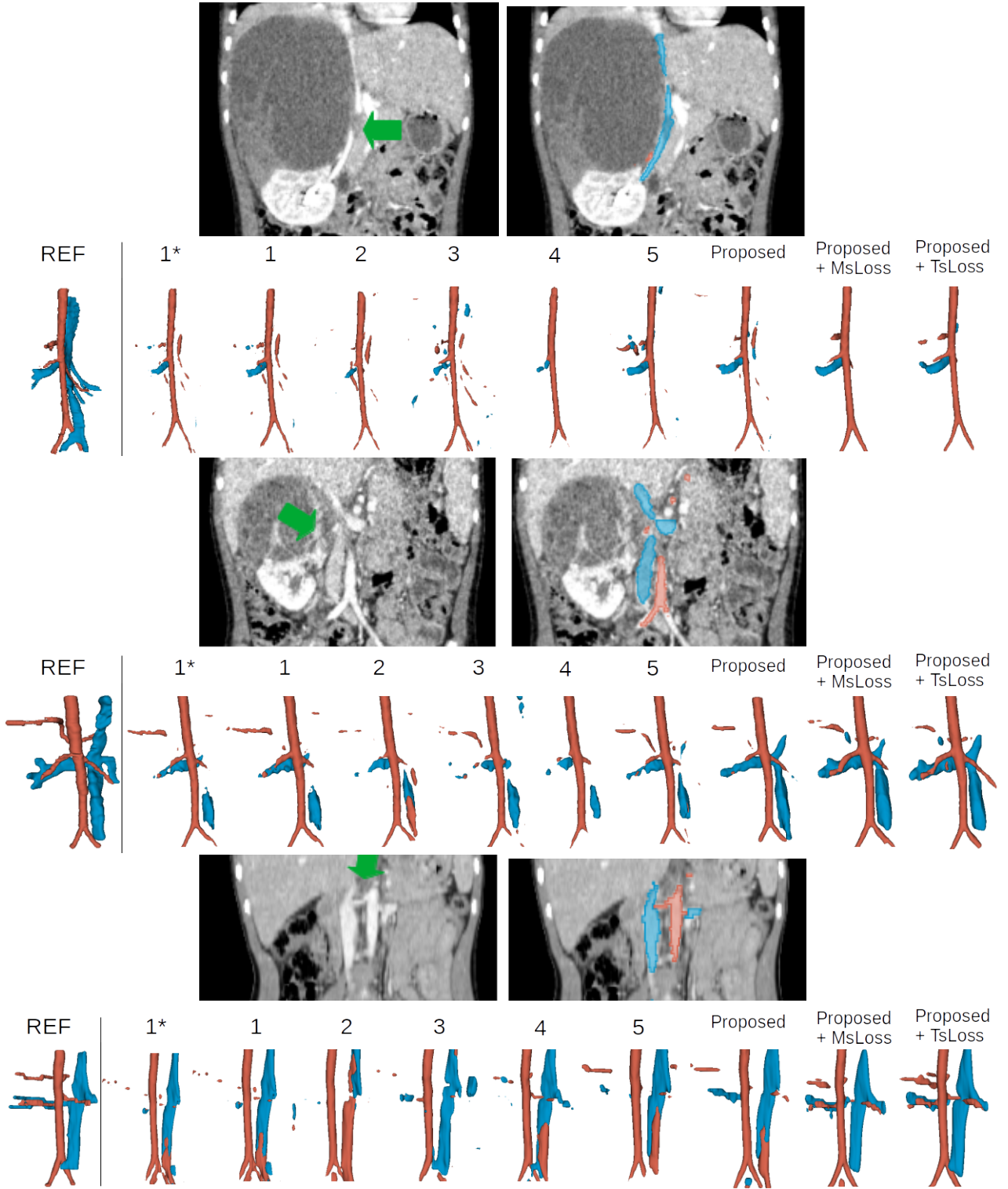


Figure 5.11: Worst (top, $DiceScore : A = 78.30; V = 15.01$), average (middle, $A = 77.94; V = 62.07$) and best (bottom, $A = 83.56; V = 84.29$) segmentation results (averaged between arteries and veins) on single patients for our method using the proposed $TsLoss$, and the results for the same patients with the other methods. For each patient we show also one coronal slice highlighting the most peculiar and difficult regions with green arrows (left: input ceCT; right: reference segmentation). The number-method correspondence is shown in Table 5.1, while the order corresponds to that shown in Table 5.2. 3D models are back-front to make renal arteries visible. Arteries in red and veins in blue. *w/o GAR.

5.4.2 Ureters segmentation

Similar considerations can also be made for ureters segmentation. However, it is good to emphasize that the database used was very limited and the test set of only 5 subjects may not be representative.

Quantitative results are shown in Table 5.3. Here the most performing state-of-the-art method, from a spatial overlap and morphological point of view, is DenseBiasedU-Net [51] (M3). This may be due to the very small thickness of the ureters (comparable to renal blood vessels, see Chapter 2), whose information is better propagated thanks to the use of the dense connections. Nevertheless, memory usage is already up to the limit with such a network and adding another loss term such as the one we propose is not possible. Moreover, the results of DenseBiasedU-Net have a high number of false positives that produce significant errors as we can see from the high 95HD. The other networks show similar behaviors to those for blood vessel segmentation, with nnU-Net [61] (M5) performing better when not considering the post-processing step (namely GAR) of the Deep Distance Transform [139] method (M1). When applying our oversampling method to nnU-Net we get worse results for overlapping measures (lower Dice score and combination of Precision and Recall), while better for spatial and morphological measures (lower Hausdorff distance and morphological similarity). The use of *MsLoss* improves all the measures, with a particular decrease in false negatives. The use of *FvLoss* in combination with *MsLoss* worsens these results, and this may be due to an inappropriate choice of parameters or to the not tubular shape of the renal calyces (i.e. the beginning of excretory pathways that is usually segmented as ureters, as in Figure 5.3). Finally, it is important to note that, unlike what was observed for blood vessels, there are generally few differences in quantitative results among the different techniques. The high standard deviation for all measures as well as a very limited test set, as mentioned earlier, make it difficult to draw conclusions with confidence.

New experiments will be performed once more data will be collected, and a semantic division between renal calyces and ureters should be performed in the manual reference segmentation.

Table 5.3: Results on 5 patients using bbox patches $96 \times 160 \times 160$ obtained with the methods of Table 5.1 and with our proposed method for ureters (U) segmentation. *without GAR. **without HRA and DPA. Mean and standard deviation of the results are provided.

Method	Oversampling	S	Dice Score [%] (\uparrow)	Precision [%] (\uparrow)	Recall [%] (\uparrow)	95HD [mm] (\downarrow)	$\Delta\Lambda$ (\downarrow)
M1 - Deep Distance Transform* [139]	Proposed <i>MinPix</i>	U	54.47 (28.49)	79.94 (12.03)	49.82 (29.72)	18.45 (27.12)	23.27 (4.46)
M1 - Deep Distance Transform [139]	Proposed <i>MinPix</i>	U	55.50 (28.17)	83.83 (9.98)	48.50 (27.93)	17.71 (25.62)	22.95 (4.36)
M2 - Deep Distance Transform [90, 139]	Proposed <i>MinPix</i>	U	46.69 (28.77)	81.22 (13.56)	43.03 (34.25)	18.01 (24.62)	24.69 (6.28)
M3 - DenseBiasedU-Net** [51]	Proposed <i>MinPix</i>	U	57.31 (28.14)	70.25 (8.96)	57.45 (31.05)	23.42 (24.87)	21.45 (3.91)
M4 - Kid-Net [128]	Random and dynamic weighting	U	50.38 (26.66)	79.97 (14.84)	46.07 (31.99)	17.71 (24.35)	23.95 (5.97)
M5 - nnU-Net [61]	Foreground in 33.3% of mini-batch	U	54.55 (24.82)	78.95 (13.20)	49.28 (29.35)	19.02 (24.76)	23.21 (4.97)
Proposed 1: Deep Sup. Dice + CE	Proposed <i>MinPix</i>	U	53.58 (24.03)	80.55 (10.40)	45.15 (29.95)	18.17 (24.29)	22.66 (4.32)
Proposed 2: Deep Sup. Dice + CE + MsLoss	Proposed <i>MinPix</i>	U	59.51 (25.85)	79.89 (6.53)	54.41 (26.85)	8.15 (12.47)	20.90 (5.65)
Proposed 3: Deep Sup. Dice + CE + TsLoss	Proposed <i>MinPix</i>	U	53.30 (29.49)	84.03 (12.04)	49.64 (33.79)	17.09 (25.78)	21.15 (3.02)

Differences are barely visible also for qualitative results. In Figure 5.12, worst, average and best results from each method are displayed. Some ceCT images have a high contrast heterogeneity in these structures, particularly in the case of biphasic injection, such as for the patient shown on the top of Figure 5.12. Furthermore, the low number of voxels of the ureters makes overlap measurements not very reliable, as we can see from the last two patients of Figure 5.12 (predictions almost complete but Dice score under 80%). Eventually, thanks to this figure it is easier to understand the fineness of these structures on the tubular section and the presence of the renal calyces discussed above, which both make segmentations even more complicated.

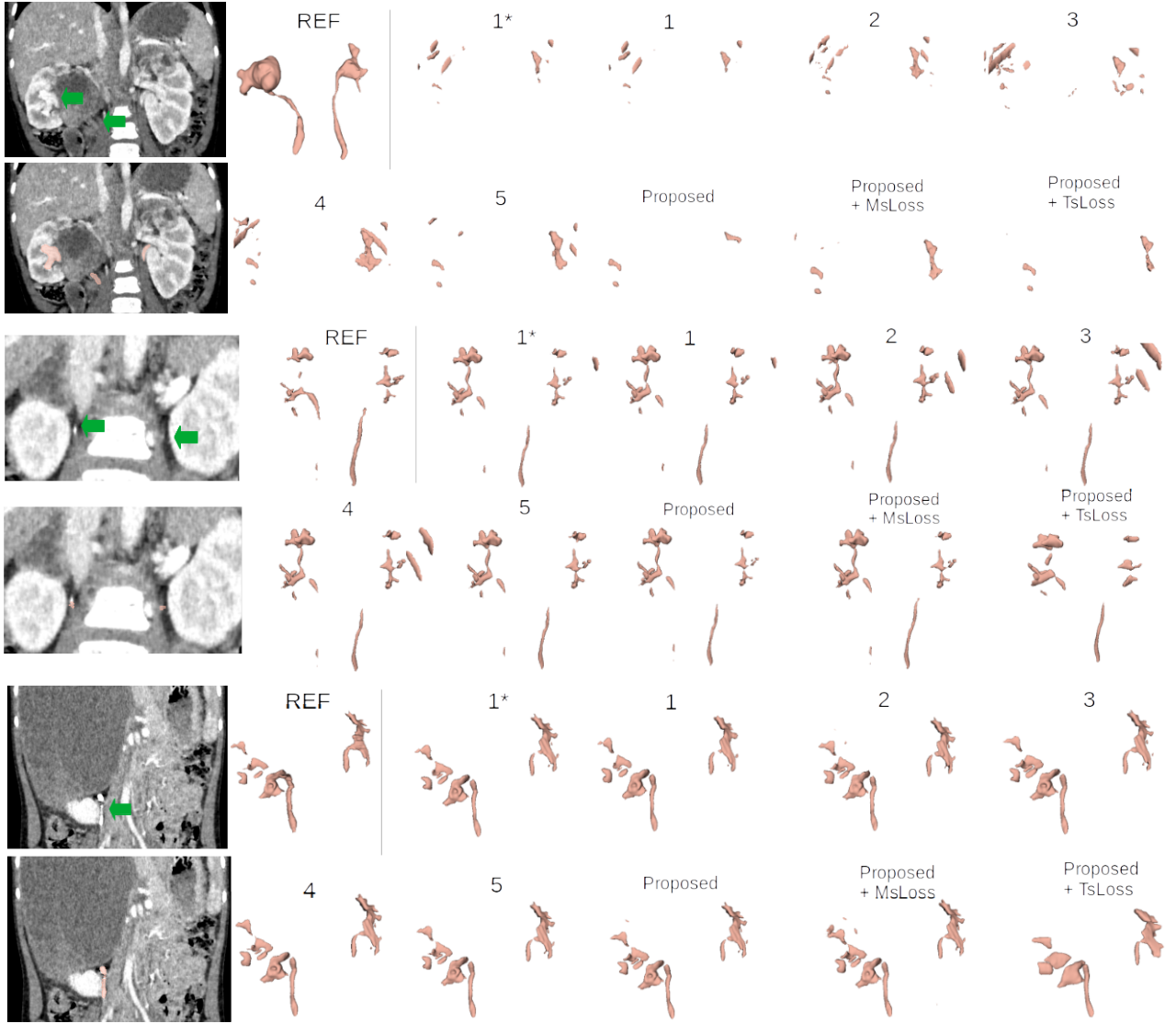


Figure 5.12: Worst (top, $DiceScore : U = 14.92$), average (middle, $U = 67.80$) and best (bottom, $U = 79.05$) segmentation results on ureters (in color pink) on single patients for our method using the proposed $MsLoss$, and the results for the same patients with the other methods. For each patient we show also one coronal slice highlighting the most peculiar and difficult regions with green arrows (left: input ceCT; right: reference segmentation). The number-method correspondence is shown in Table 5.1, while the order corresponds to that shown in Table 5.3. The differences are barely visible and do not result in major differences from the point of view of subsequent manual correction. *w/o GAR.

5.5 Conclusions

In this chapter, we proposed for the first time an assessment and comparison of state-of-the-art methods for segmentation of renal tubular structures (arteries, veins and ureters) in ceCT images of pathological and pediatric patients.

We focused on the works centered on these images and structures, but others that we found interesting were also analyzed. Probably some interesting works have not been analyzed, particularly among the methods that we called rule-based or among *non-deep* machine learning approaches. For that reason, the assessment may not be complete, but preliminary tests done with these methods resulted in great difficulty in segmenting the renal tubular structures in pediatric ceCT images acquired on arteriovenous phase.

Also in terms of comparison, we probably could have examined other methods among those presented, but the lack of codes and the few details available in the articles did not allow us to reproduce them with confidence. For the chosen methods that have no code available online, I believe that the implementations are correct but small errors could always be present. We would also like to mention that other tests with other loss functions such as general Dice or the use of a distance map to weigh voxels were performed, but the results did not show significant differences with those of the chosen methods that also exploited these techniques.

We proposed also the use of a loss function designed from the so-called vesselness function to improve state-of-the-art results. This loss function is based on the comparison of eigenvalues of the Hessian matrix of segmentation masks and Frangi's vesselness enhancement on target voxels in a multi-scale deep supervision way. The combination of this tubular structures loss function with voxel-wise loss functions allowed us to overcome some problems of the latter, such as the difficulty in correctly optimizing tubular structures with elongated shape, intra-scale changes and inter-anatomy variation. The results demonstrated great improvements from a morphological point of view, with segmentation results showing fewer interruptions, at the expense of a slight increase in false positives. This confirms that the use of voxel-wise loss functions and overlapping measures is not sufficient for the evaluation of such structures. The use of the second loss term, related to Frangi's vesselness, appears to be of no benefit in cases where the structure has non-tubular regions. In addition, several hyperparameters are introduced with the use of this loss function, therefore an automation of the choice of these parameters is planned.

The results of applying the method proposed here with the combined use of real ceCT and synthetic CT images, generated via the method proposed in the previous chapter, are shown in the next chapter. However, it is important to anticipate that the use of the tubular structures loss function already partly succeeds in tackling the heterogeneity in contrast intensity of ceCT images.

Chapter 6

Application of anatomical digital twins for renal cancer surgery

In this chapter, we show on real cases the clinical benefits of the 3D anatomical model, built from the approaches proposed in previous chapters via the use of a software tool designed for doctors. In Section 6.1 we present the software tool developed as 3DSlicer [37] plug-in for the IMAG2 lab of Necker hospital. Then, in Section 6.2, we discuss the advantages of using 3D anatomical digital twins for pre-operative planning and per-operative guidance, showing some interesting clinical cases and an on-going clinical study to further evaluate these benefits.

6.1 Preparation of anatomical models via a 3DSlicer plug-in

6.1.1 From ceCT scan to 3D volume: the “Renal Anatomy Segmentation For ceCT” module for 3DSlicer

In order to use the proposed methods in a real clinical setting, a software tool for clinicians was developed as a plug-in for the 3DSlicer¹ [37] open source software. To the best of our knowledge, this is the first plug-in specifically dedicated to renal anatomy segmentation of pediatric patients with kidney tumors. The goal of our tool is to speed-up the annotation of the renal anatomy from ceCT scans, and to reduce the manual interactions. Starting from the excellent results already obtained with the automatic segmentation methods, the medical experts have only to refine these results. Our plug-in is developed as a module in 3DSlicer, named “Renal Anatomy Segmentation For ceCT” module. An overview of the module is shown in Figure 6.1.

The module provides 5 sections:

- **Help & Acknowledgment.** Here there is a brief explanation of the module, the main contributor and entities that granted and collaborated in the creation of the plug-in.
- **Reload & Test.** This section is used by developers when creating, editing, and debugging the module.

¹version 4.11

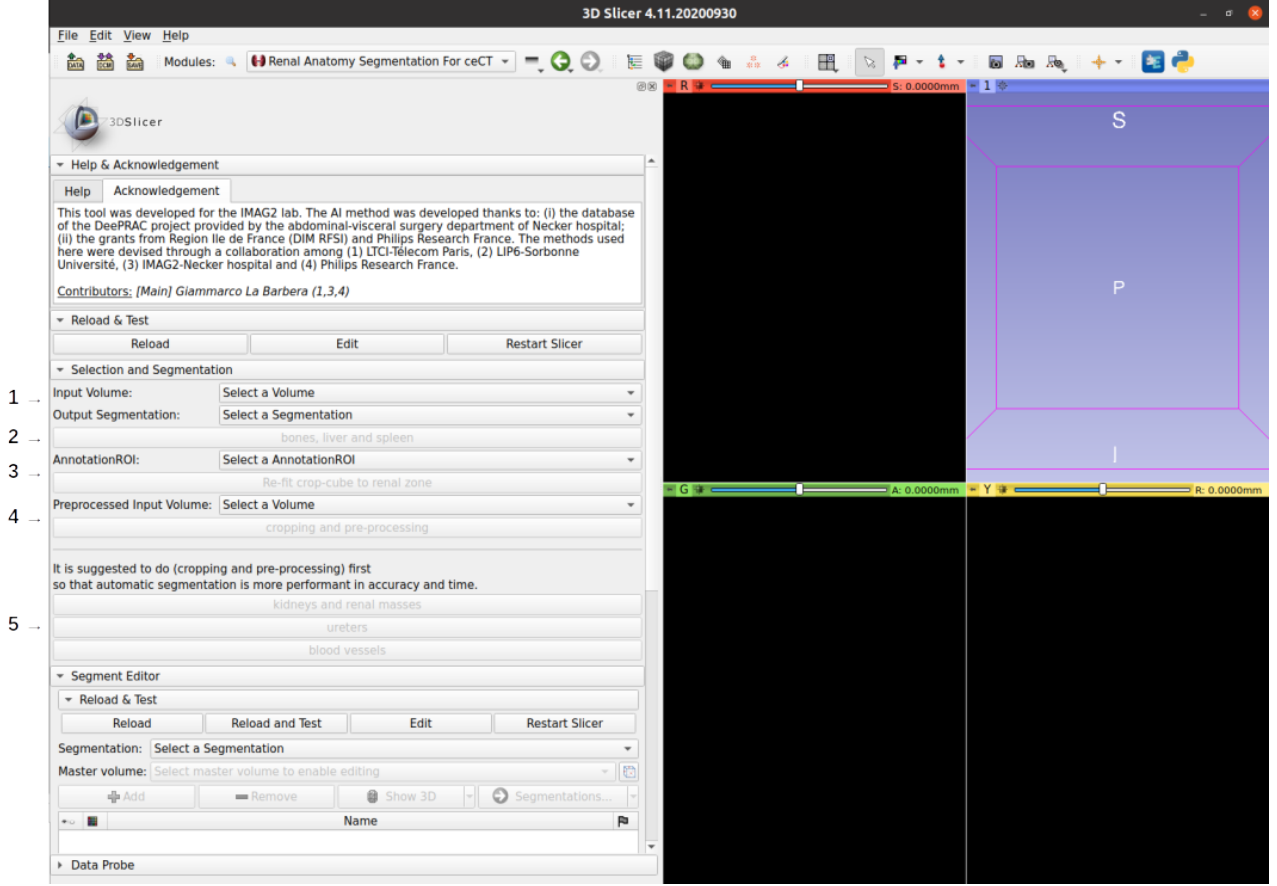


Figure 6.1: Overview of our “Renal Anatomy Segmentation For ceCT” module for 3DSlicer [37]. The frames labeled 1 to 5 correspond to the steps of the selection and automatic segmentation: 1. input and output selection; 2. bones, liver and spleen automatic segmentation; 3. initialization of the annotation ROI; 4. selection of the preprocessed input volume to speed-up point 5; 5. automatic segmentation of kidneys, renal masses, ureters, arteries and veins. See text for details.

- **Selection and Segmentation.** This is the main section of the module, which is described in detail in the next paragraph.
- **Segment Editor.** This is the original 3DSlicer module for manual and semi-automatic segmentation that is linked to this section to allow the user to be able to refine results provided by the automatic segmentation method in the same interface.
- **Data Probe.** Default section of 3DSlicer that allows the arrow-pointer interaction with the displayed image, providing data of where the indicator is located, such as position in the 3D volume, HU or label value, name of segmentation structure and more.

The **Selection and Segmentation** section is decomposed into 5 different tasks (see Figure 6.1) to facilitate the user-interaction and to speed-up the inference phase for the automatic segmentation.

- 1 Once the medical image data (i.e. the ceCT scan) is loaded into 3DSlicer, the user can select it as input volume. Then, in order to activate the successive steps, the user has to create an output segmentation volume, where the results can be recorded. When both volumes are selected, the other buttons become active.

- 2** This button allows for an automatic segmentation of bones, liver and spleen, in order to have a 3D model as complete as possible. These structures are usually not affected by the renal tumor and after a pose and size homogenization to an adult reference sample, bones, liver and spleen on children result very similar to those of adults. According to this observation, we trained three different networks using our 3D nnU-Net [61] implementation on adult ceCT images obtained from two available public databases: *CT-ORG* [115] from the Cancer Imaging Archive (TCIA) [24], with 135 ceCT images of pathology-free adult bodies where lungs, bones, liver, kidneys and bladder are labeled; and *CT-Abdomen* from the MICCAI 2015 Abdomen Challenge [82], which includes 30 ceCT scanners of adults affected by colorectal cancer with 13 abdominal organs labeled, among which spleen and liver. We decided to focus on bones, liver and spleen because they are less affected by the spatial differences and pathologies. We trained three different networks: a first one on bones using ceCT images from *CT-ORG* (115 for training, 20 for test), a second one on spleen using ceCT images from *CT-Abdomen* (25 for training, 5 for test) and a third one on liver using images from both database (140 for training, 25 for test). Results on adult test sets are shown in Table 6.1.

Table 6.1: Results (mean and standard deviation) on adults test sets (20 patients for bones, 5 for spleen and 25 for liver). See text for details.

Structure	Dice Score [%] (\uparrow)	Precision [%] (\uparrow)	Recall [%] (\uparrow)	95HD [mm] (\downarrow)
Bones	87.67 (2.94)	91.35 (2.94)	84.47 (5.02)	17.97 (11.12)
Spleen	38.54 (26.76)	99.68 (0.21)	26.78 (21.93)	44.81 (15.55)
Liver	86.85 (17.96)	98.10 (0.74)	80.88 (18.80)	22.21 (24.35)

Then we proceed as explained in Section 3.5 for direct inference using a *common* pose and size but in 3D. In particular children volumes are transformed using a 3D *STN* for pose and size homogenization, then the patches are extracted and are inferred sequentially in the three pre-trained networks previously presented. The segmentation results for bones are more than satisfactory, with better results than thresholding and region growing due to hyper-contrasted arteries and veins. For liver segmentation, if the liver is not modified by a renal tumor, results are comparable to the ones obtained using other 3DSlicer tools such as RVXLiverSegmentation [81]. For the spleen segmentation, the already poor performance on adults, because of the limited database, resulted in poor results also on children. Please note that in the vast majority of renal tumors in children, the spleen is not invaded, so it is not a priority segmentation. Despite this clarification, we consider these results as preliminary and this step requires further investigation.

- 3** With the aim of speeding-up inference time and improving segmentation performances (reducing the amount of false positives), a 3D bounding box can be selected by the user. When the so-called “AnnotationROI” is created, the inference phase of the method proposed in Section 3.4 in its 3D version *STNpose-size* + *STNcrop* is performed. The method is applied in its entirety but is stopped when the minimum and maximum bounding box values are predicted. Since optimal performances are not yet achieved, the user can interact with the proposed ROI in order to adjust it. The user can re-fit the bounding box in the initial proposed location via a button. This step is not mandatory, and if it is not applied, then almost the whole image is used in the next steps (a bounding box of the voxels with non-zero values is used).

- 4 Once the bounding box is set, the user can create a new volume in which the cropped and pre-processed images will be saved. The pre-processing consists of the organization of the axis order to be coherent with the pre-trained network, the resampling in the common voxel space used for training and the storing of the locations of the patches via the *sliding-window* technique. This step is not mandatory, and if is not applied, the original, un-processed, images are used next.
- 5 The last part is the main automatic segmentation task. Three buttons are available to segment respectively kidney and renal masses, ureters, and blood vessels (arteries and veins). For the automatic segmentation of kidneys and renal tumors we use the weights of the network presented in Table 4.5, trained using our 3D nnU-Net [61] implementation with real ceCT images and synthetic CT images generated with the method proposed in Chapter 4. The segmentation methods presented in Chapter 3 result in lower performance in 2D. For the automatic segmentation of renal tubular structures, i.e. arteries, veins and ureters, we trained two new networks (one for blood vessels and one for collective systems) using the proposed method for tubular structures segmentation, presented in Chapter 5, using both real ceCT images and synthetic CT images generated with the method proposed in Chapter 4. Performances are summarized in Table 6.2. Each result is uploaded in the output segmentation volume created in step 1 as a segmented structure with his name (e.g. arteries) and his assigned color (e.g. red). For kidneys, tumors and ureters, if they present multiple connected regions, these are divided in different labels in the segment editor, e.g. kidneys is divided kidney1 and kidney2 (resp. left and right). Then the structures will be visible in the Segment Editor section, with which the user can interact to refine them. This step (5) can be performed even if the two previous (3 and 4) are not applied. In this case, almost the whole image is used (cropping in non-zero values) and the pre-processing is performed individually on the click of each button. This slows and reduces the performance.

Table 6.2: Best results (mean and standard deviation) obtained with the combination of the techniques presented in this thesis. For the segmentation network of kidneys and tumors the 3D nnU-Net [61] framework is used with as input for training both real ceCTs and synthetic CTs produced with the method presented in Chapter 4. The same combination of input images and the same framework, together with the oversampling method and Tubular structures Loss function proposed in Chapter 5, are used for ureters segmentation (only *MsLoss*) and blood vessels (arteries and veins) segmentation (complete *TsLoss*). The first four evaluation measures used are described in Appendix A, while the last one $\Delta\Lambda$ is the Morphological similarity Loss (*MsLoss*) described in Chapter 5.

Structures	Dice Score [%] (\uparrow)	Precision [%] (\uparrow)	Recall [%] (\uparrow)	95HD [mm] (\downarrow)	$\Delta\Lambda$ (\downarrow)
Kidneys	90.84 (3.56)	89.46 (4.20)	92.38 (4.17)	4.62 (4.39)	-
Tumors	88.39 (11.71)	92.17 (9.40)	86.80 (14.96)	6.46 (6.54)	-
Ureters	61.66 (24.29)	77.89 (7.49)	57.61 (27.44)	8.17 (13.13)	20.17 (6.10)
Arteries	75.34 (4.90)	86.41 (6.46)	66.47 (7.17)	11.32 (8.96)	19.73 (1.50)
Veins	61.47 (18.92)	82.73 (9.56)	50.99 (20.01)	16.47 (15.05)	19.13 (3.24)

These steps are illustrated via screenshots taken during its use in Appendix F. The cropping and segmentation algorithms are implemented through the use of Docker [98], an open source containerization platform which enables applications to run quickly and reliably from one computing environment to another.

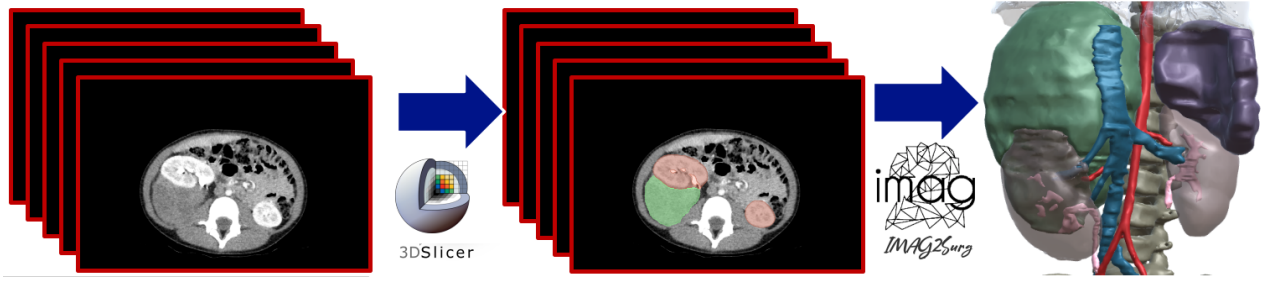


Figure 6.2: Summary of the essential steps of the proposed method from ceCT scans to 3D digital twins via the use of our 3DSlicer [37] plug-in, presented in Section 6.1.1, and the software tool specifically designed at the IMAG2 lab of Necker hospital for visualization and interaction, presented in Section 6.1.2.

6.1.2 From 3DSlicer to the operating room: an anatomical digital twin to help surgeons

Once the manual corrections are completed, the complete segmentation is shown to a member of the Necker pediatric radiology team² for a first evaluation. If it is necessary, further corrections are made and the segmentation is also subsequently finally reviewed by a member of the Necker pediatric abdominal-surgery team³. Then, the 3D output segmentation volume is exported as object and processed via the software Blender [56] for rendering and texturing. The final anatomical digital twin is uploaded in a software tool specifically designed at the IMAG2 lab of Necker hospital, which allows for better visualization and more intuitive interaction than 3DSlicer, making its use easier for both clinicians and patients. This 3D reconstruction is in fact used to show in a more clearly way the case to the patients (and their parents), but more importantly for improving the pre-operative planning as is presented in the next section. A summary of the essential steps with an example of a complete digital twin is shown in Figure 6.2 (also in Figure 1.1 in the Introduction). Eventually, the digital twin is displayed in the operating room during the whole operation, as shown in Figure 6.3.

During the operation, a surgical resident or a member of the IMAG2 team are in charge of orienting and manipulating the 3D model according to the surgeons' indications.

We would like to emphasize that the number of manual corrections required is usually very low, reducing the total segmentation time of 8 hours compared to that of a totally manual segmentation [21] (if the user is not expert in the use of 3DSlicer tools) and of more than 2 hours compared to a semi-automatic one (expert user in 3DSlicer tools). The total automatic segmentation time for all the 5 steps is about 6 minutes using a GPU with 11 GB of VRAM. The time needed by two different 3DSlicer experts is shown in Table 6.3, and is the total time averaged for the segmentation of three subjects.

6.2 Advantages of 3D models in pre- and per-operative planning

As discussed in the Introduction, the relationships of the tumor to kidneys, renal vessels and excretory systems must be perfectly known, in order to decide for the type of surgery to fit

²service of Pediatric Radiology of Hôpital Necker Enfants-Malades, head of service: Pr Natalie Boddaert.

³service of Pediatric Visceral, Urological and Transplant Surgery of Hôpital Necker Enfants-Malades, head of service: Pr Sabine Sarnacki.



Figure 6.3: Picture taken during a surgical operation of a pediatric renal tumor at Necker hospital. The arrow highlights the 3D anatomical digital twin displayed next to the images acquired by the operating camera.

Table 6.3: Total segmentation time (mean and standard deviation) needed by two different 3DSlicer experts, averaged for the segmentation of three subjects. With the use of the plug-in, 6 minutes are necessary to complete the 5 steps of the automatic segmentation. *This time is the one reported in [21] for 14 patients without bones, liver and spleen segmentations.

User	Use of the plug-in	Total segmentation time
Non-3DSlicer experts [21]		9h (6h)*
3DSlicer expert 1		4 h (1 h)
3DSlicer expert 2		3 h (30 min)
3DSlicer expert 1	✓	1 h (15 min)
3DSlicer expert 2	✓	1 h (10 min)

to the criteria of the Umbrella SIOP protocol [12]. However, CT scans are a sequence of 2D slices and cognitive volume reconstruction can be challenging for surgeons [133]. The patient-specific 3D virtual model not only allows for an easier visualization of the renal anatomy, but it can also improve the surgical planning and intraoperative guidance [59, 100, 109, 133]. This is even more important in pediatric patients, yet few works validate this claim [140, 60] and anatomical digital twins are still not currently routinely used.

6.2.1 Analysis of two interesting clinical cases

We report here two relevant clinical cases of pediatric nephroblastoma faced by the team of Pr Sabine Sarnacki, in which the 3D model proved to be more effective than the visual inspection of series of 2D images.

The first case is a 20-months-old child who had a bilateral nephroblastoma. The pre-operative bi-phasic ceCT scan (after 12 weeks of chemotherapy) still showed multiple tumor

masses in both kidneys. Some slices are shown in Figure 6.4.

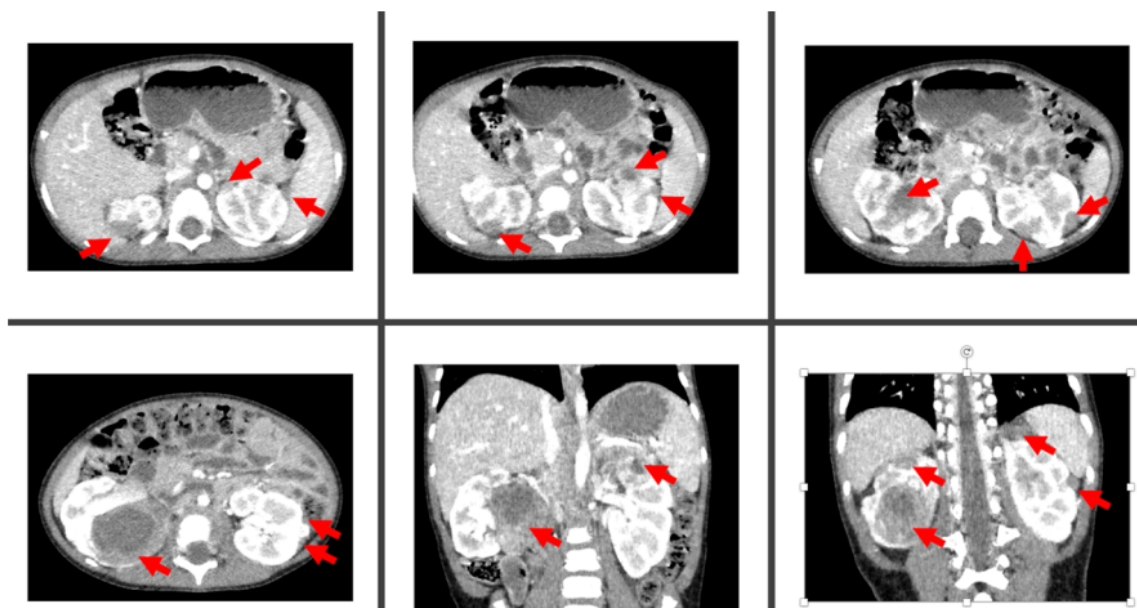


Figure 6.4: Pre-operative bi-phasic ceCT scan of the first case. Renal tumors are highlighted by arrows.

On the right hand side, he had a large central tumor of 6 cm in diameter, associated with several nodules of diameter ranging from 5 to 10 mm, and on the left hand side, a big tumor in the upper pole of 2.6 cm in diameter, associated with several nodules from 3 to 8 mm in diameter. From the analysis of this ceCT scan, the proposed surgical planning was partial nephrectomy (Nephron-Sparing Surgery) for the left kidney and radical nephrectomy (RN) for the right kidney considering the central location of the main tumor on this side. Subsequently, the 3D anatomical digital twin was developed from the ceCT image, following the steps described in Section 6.1. The result is shown in Figure 6.5. It is important to point out that in this case the 3D model was manually corrected by adding information obtained from MR images acquired out-of-protocol given the complexity of the case.

A second assessment using the 3D virtual reconstruction was performed, where the plane of dissection of the central mass of the right kidney with the pelvis was clearer and allow to attempt a partial nephrectomy (NSS) also on the right side (removing and dissecting the central mass from the excretory system). The new planning was chosen and adopted during the surgical procedure.

The intraoperative findings corresponded perfectly with the 3D reconstruction, as visible in Figure 6.6, and the per-operative guidance (detailed in Section 6.1) was reported by the surgeons to ease the masses localization and identification.

The operation was successfully completed, following the new surgical plan with NSS for both kidneys, prepared thanks to the 3D model. The success of this operation is also to be given to the ability of the 3D digital twin to fuse information from different modalities (such as ceCT and MR as here).

This case study led to an abstract accepted at Congrès annuel de la Société Française de Chirurgie Pédiatrique (SFCP) 2021 [SFCP-21].

The second case we report is a 6-years-old child who had a unilateral nephroblastoma on the right kidney, where the decision trend was opposite to that of the first reported case. From the pre-operative arteriovenous ceCT scan, despite the central location of the tumor

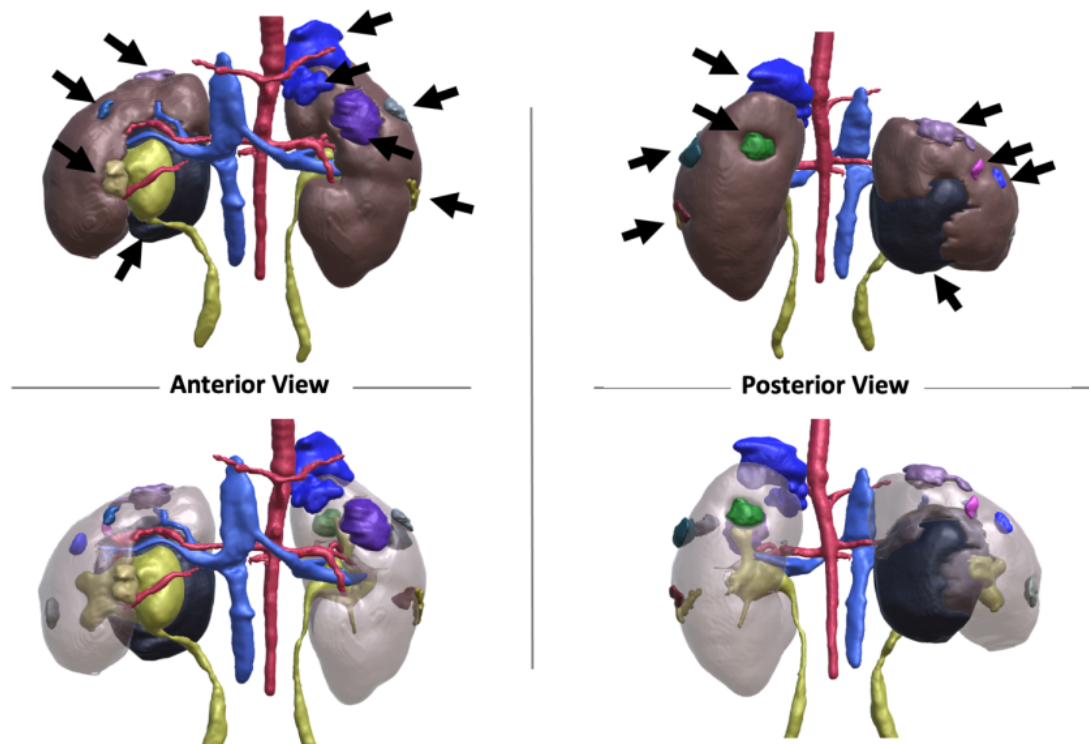


Figure 6.5: 3D anatomical digital twin of the first case. The structures represented are kidneys (in brown), arteries (red), veins (blue), ureters (yellow), tumors (other colors, highlighted by arrows). Left: Anterior view. Right: Posterior view.

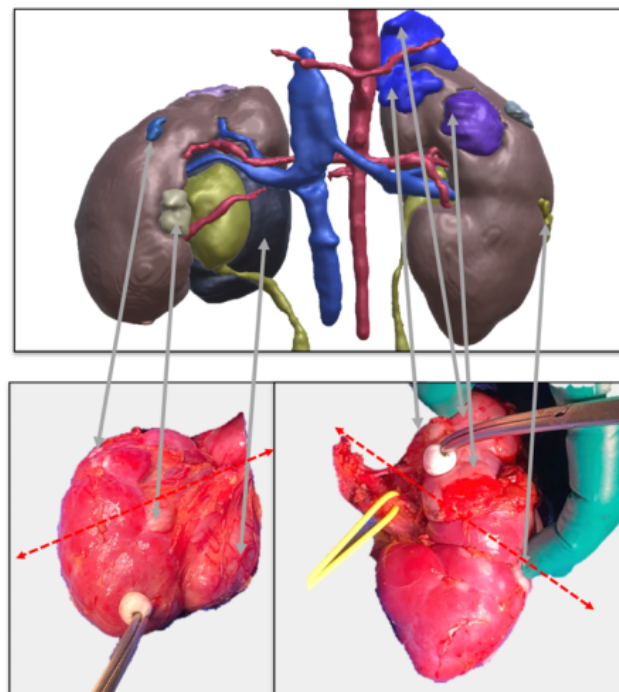


Figure 6.6: Correspondence in the first case between 3D anatomical model (top) and the intraoperative findings (bottom).

and its size, the surgical team wondered whether it was possible to proceed as in the previous case with a NSS. The ceCT images did not allow to make a choice with high confidence, and a solution would have been to plan a partial nephrectomy, moving on to a RN if that was impossible, with all the associated risks. Some slices are shown in Figure 6.7.

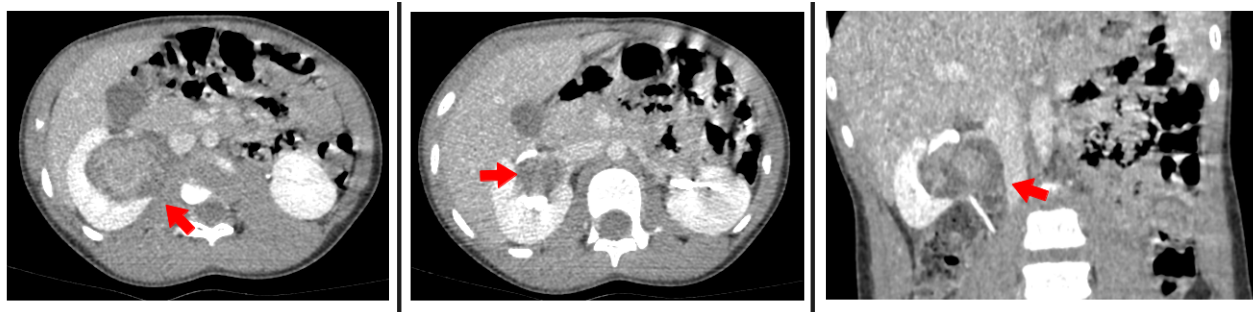


Figure 6.7: Pre-operative bi-phasic ceCT scan of the second case. Renal tumor highlighted by arrows.

Once the 3D anatomical digital twin was developed a second assessment was performed. Here, the surgical team found that: (i) the amount of the spare renal parenchyma would have not met the criteria of the Umbrella SIOP protocol [12]; (ii) several blood vessels and ureters passed in close proximity to and within the tumor. These conditions compel the performance of a radical nephrectomy. The 3D reconstruction is shown in Figure 6.8.

Again, the operation was successfully completed, confirming that a partial nephrectomy would not have been possible. This was also confirmed by the pathology report (this is a study to know whether the renal sinus had been infected by the tumor, meaning that the kidney is entirely compromised). In this case, the 3D model was helpful in making a final decision more confidently, and in being able to plan precisely the surgery from the beginning, avoiding any possible risk. A partial nephrectomy could also have led to a microscopic incomplete resection, requiring a post-operative radiotherapy to avoid recurrence as recommended by Umbrella SIOP protocol [12].

6.2.2 Retrospective on-going study on 3D model vs. 2D imaging

To further validate the benefits of the 3D anatomical model versus using only 2D structural imaging, a retrospective study was designed as follows.

1. First, among all the available patients (see Chapter 2), we selected 20 pediatric patients with renal tumors who have undergone surgery at Necker hospital, with a tumor stage eligible for a NSS (no ganglions, thrombus or calyces involved). In particular we selected 9 patients in which a NSS was performed and 11 in which a RN was then conducted (both easy and difficult cases). All patients underwent preoperative ceCT imaging and for some of them the complete 3D models were already obtained and used for planning and guidance. For the patients who did not have a complete 3D model, we proceed as explained in Section 6.1. All images were anonymized and the models transferred to a tablet.
2. Four experts in nephroblastoma surgery and one radiologist from four different centers were already solicited to participate to the study. In contrast to what was stated in the previous section for surgeons, radiologists are more accustomed to simulate three

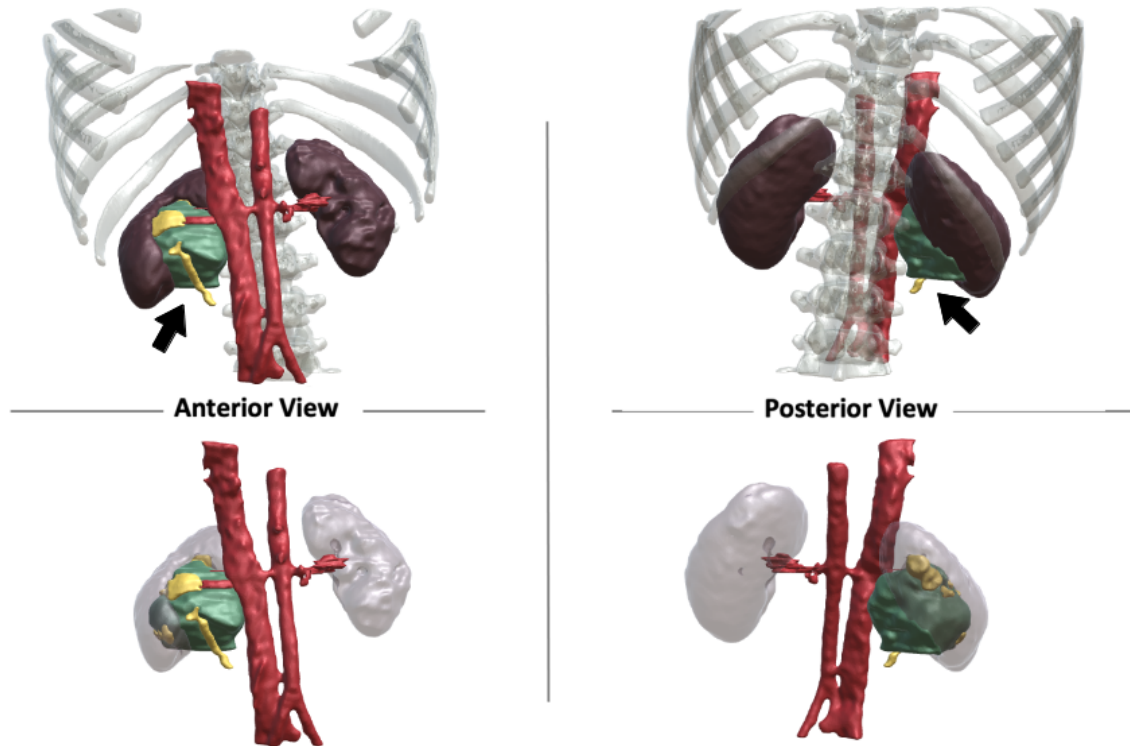


Figure 6.8: 3D anatomical digital twin of the second case. The structures represented are kidneys (in brown), blood vessels (red), ureters (yellow), tumors (green, highlighted by arrows). Left: Anterior view. Right: Posterior view.

dimensional anatomy in their mind. Therefore, at least two other radiologists will be involved in the study.

3. Each physician analyzes at a first time every pediatric case using only the 3D anatomical model and proposes a first surgical planning. Decision-making reasons and time are recorded. We start from the 3D model because we hypothesize a longer decision time on the series of 2D slices that would help the physician remember the patient.
4. After one month from the previous step every patient is analyzed a second time using instead the 2D structural imaging (i.e. same ceCT scans from which the 3D models were built). The delay is set in order to try to make the physician forgetting the surgical plans proposed in the previous step, moreover patients are presented in a different order. Thus a second surgical planning is proposed, and the new decision-making reasons and time are also recorded.
5. The new two plannings proposed by the experts are compared with the procedure previously performed for the surgery.
6. The new two plannings are also compared with the post-operative anatomical pathology of the removed kidney, either entirely (for RN) or partially (for NSS). As previously explained, a negative anatomical pathology response confirms whether the choice to proceed with NSS instead of radical nephrectomy is correct, and will then be used as a reference.

The questions we want to answer are the following ones.

- Non-inferiority of the 3D models compared to the series of 2D slices for decision-making reasoning. The 3D digital twin is built from the ceCT images, therefore no more information is available for decision-making (only a better visualization and an easier navigation).
- Inferiority of the 2D imaging compared to the 3D model in decision-making timing.
- Number of cases where an NSS could have performed instead of RN thanks to the use of the 3D model for planning.
- Number of cases in which a RN would be chosen because of risks that could only be identified in the 3D model.
- Reduction of inter-subject variability in surgical plan assessment with 3D models compared with 2D imaging. In particular here we can also use cases where the 3D model had already been used for that choice.
- Existence of intra-subject variability in surgical plan assessment with 3D models compared with 2D imaging. We want to confirm what was assessed in [133] for difficulties in cognitive volume reconstruction by surgeons, and examine if also radiologists in some particular clinical cases may encounter these difficulties.

This study is still ongoing.

6.3 Conclusion

In this chapter we have shown how the approaches proposed in this thesis are used in a real clinical setting.

The 3DSlicer plug-in we developed allows to exploit these methods to automatize large part of the segmentation process, reducing the 3D model creation time and the manual interaction required from clinicians. However, some parts of this plug-in still need improvements. Moreover, a proper evaluation protocol should be conceived to prove the system effectiveness in clinical practice.

Furthermore, after introducing how the final 3D model is used for pre-operative planning and per-operative guidance, two relevant cases were analyzed where the use of the anatomical digital twin enabled more effective and confident decisions in both of these steps. In one of two cases, it was also shown that information from two different acquisition modalities (e.g., ceCT and MR) could be coupled into a single 3D digital twin.

The advantages of using such 3D models for choice over radical or partial nephrectomy and over laparoscopy or laparotomy surgery need wider validation in the literature, particularly to support their relevance in pediatric cases. An on-going study of our own to that end was introduced here.

Chapter 7

Conclusions and perspectives

The final goal of this PhD thesis was to create individual 3D anatomical models from pediatric abdominal-visceral ceCT scanners with renal tumors leveraging deep learning techniques. This model allows for an easier visualization of the renal anatomy, improving surgical planning in order to save as much functional kidney tissue as possible and intraoperative guidance [60, 140] allowing minimally invasive laparoscopic procedures [109]. These 3D digital twins are based on image segmentation which was usually performed manually by clinicians via softwares such as 3DSlicer [37]. The use of deep learning approaches aims to automatize this process, speeding up model creation and reducing manual-interaction required from the physician.

Nevertheless, the analysis of our pediatric and pathological abdominal-visceral ceCT images raise several difficulties, as detailed in Chapter 1.

7.1 Conclusions

In order to reach our goals, four contributions have been made (presented in Section 1.2 and summarized in Figure 1.5). Their achievements and short-term perspectives are discussed in this chapter.

Segmentation of kidneys and renal tumors. The segmentation of the main structures of the 3D renal model on pediatric images, namely the kidneys and the tumors, can be based on the extensive literature available on adults, thanks also to the MICCAI challenges KiTS19 [53] and KiTS21 [52]. The best-performing methods are based on the use of U-Net [117] networks, particularly on the well-known nnU-Net [61] (winner in the 2019 challenge). The in-depth analysis, with an implementation of its pipeline from scratch, allowed an understanding of the techniques and steps needed to achieve high segmentation performance in medical images. We tested nnU-Net in both 2D and 3D versions on our pediatric images, and we can conclude that: (i) for the kidney segmentation, despite the changes made by the tumor to its parenchyma, the use of a few subjects (for training and validation) already leads to satisfactory performance (close to a Dice score of 90%); (ii) for the tumor segmentation, more than twice as many subjects are needed to achieve such performance together with the use of a 3D U-Net. The latter, however, shows lower performances than the 2D U-Net as the number of subjects decreases. Personally, I found this point very interesting: the main difference lies in the extraction of information, which in 2D networks is done on 2D slices of the 3D volume while in 3D networks, due to computational limitations, on 3D patches (three-dimensional portions

extracted from the ceCT volume). The latter exploits information in the third dimension, which is significant in medical imaging. However, the limited size of the patches cannot always cover the entire target structure, and the network has more difficulty extracting knowledge related to the shape, size and pose, as well as relation with other structures. Working with slices instead, the network learns this information more easily but limited to the cross section (e.g. axial) used for training. For this reason, the network also performs well with a smaller amount of samples but cannot improve its performance much. These problems are even more pronounced when working with a limited yet heterogeneous database in size and pose, such as the Necker PRAC database.

In order to cover all possible transformations that are not present in the training set, data augmentation techniques are often used. For example, the high performance of nnU-Net is based on a strong use of it. However, we believe that it is really difficult to cover all possible size and pose transformations with such a technique, which is also very time-consuming and memory-intensive. We have proved that the use of automatic homogenization techniques through the use of Spatial Transformer Networks [65] in order to reduce such variability are more efficient, in terms of performance, time and memory. In addition, STNs can also be used to zoom or crop on the region of interest, increasing even more segmentation performance or reducing time and memory requirements, while maintaining high performance, respectively. It is important to emphasize that the use of a CNN for localization is already found in literature [61, 78, 156] but, to the best of our knowledge, no network has the ability to automatically zoom or crop as proposed here.

These positive conclusions to our first question concern only the implementation in 2D. In fact, in order to use the proposed techniques in 3D, the entire ceCT volume has to be provided to the STNs. Memory limitations can be overcome by resampling at a smaller size, as homogenization of the whole body as well as ROI detection does not need high details. Nevertheless, operating in this way, the number of samples available is really small and the training falls into the overfitting problem. The most widely used technique to overcome this issue is precisely data augmentation. This means that in a 3D scenario, data augmentation technique are still better than an homogenization approach. Moreover, even with a collection of images such that we can train the STNs correctly, we should have a GPU powerful enough to take as input the extracted 3D ROIs. Working with our pediatric patients, I do not believe this is possible given the extent of tumors in some subjects, but probably on other databases this method could be applied. For the purpose of fair comparison, if such powerful GPU is available, other methods than convolutional autoencoders should be tested, such as Vision Transformers [31, 49] and ConvNeXt [88]. But the question to ask here would be: given the already high performance obtained with 3D nnU-Net, is it really necessary to use this high carbon-demanding computational power? Examining patient by patient in Figure 7.1 we note that only one subject has both Dice score and Hausdorff distance outside the limits of “good” performance (that we set to 80% and 20 *mm* respectively). If we also take into account the inter-variability of manual segmentation among clinicians, the ease of correction of the automatic segmentation through dedicated tools such as 3DSlicer [37] and the difficulty of clinicians in accepting 100% automatic segmentations, I would provide a negative answer to this question.

This reasoning is only solid if we have a pediatric database available with a certain number of patients, such as the one we managed to gather at Necker hospital of Paris. Less important centers do not have this possibility, and using less data as we have examined leads to a

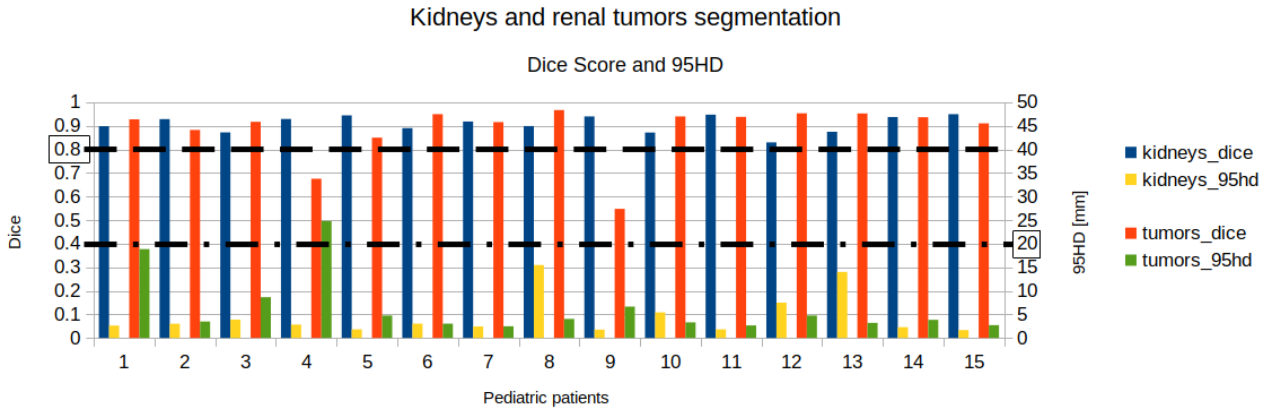


Figure 7.1: Dice score (left y-axis) and 95th percentile of Hausdorff Distance (right y-axis) of segmentation results on each of the 15 patients of the test set of the Necker PRAC database using nnU-Net 3D trained with 65 patients. The bold line are “good” performance limits set by us.

drop in performance. Yet more and more often the weights of pre-trained nets are shared in inter-center projects [123] or available online, as are those of nnU-Net winner of the KiTS19 challenge. However, the differences in pose and size between adults and children do not enable their direct use or an efficient transfer learning, despite a strong use of data augmentation, as we experimented. The use of the proposed STN to homogenize size and pose allows us to improve the segmentation of the kidneys in direct inference on 15 pediatric subjects (but not of the tumors, whose segmentation performance remains really low). While managing to recover another 25 patients for training, the combination of STN with fine-tuning on adult weights leads to results on par with training with 65 pediatric subjects. However even these experiments are only with 2D networks. In this scenario, being able to experiment with this method using 3D networks would be useful for the research community and the use of more powerful GPUs would be justified.

Cross-domain CT image translation using CycleGAN with anatomical constraints.

Some anatomical structures such as renal tumors, blood vessels and ureters can be challenging to segment in abdominal ceCT images also due to the variability in contrast medium diffusion. Inspired by some recent works [119, 125, 158], we wanted to leverage the use of both ceCT and contrast-free CT images to improve the segmentation performances. Due to the major presence of only one modality per subject in the Necker PRAC database (to limit radiation dose), namely ceCT, we propose to compensate for the lack of the other modality using generative models. The only 10 subjects who have undergone also on a contrast-free CT are not sufficient to train with success a network, and as a consequence we conducted the experiments for ceCT-CT translation task using two unpaired datasets of non-pathological adult patients.

Tests on unsupervised translation state-of-the-art methods showed two interesting points. First, the original CycleGAN [157] with Res-Net as generator network and PatchGAN as discriminator mechanism still performs better than other recent proposed methods [2, 68, 72] which probably require high number of samples and large memory to reach satisfactory results. Then, also the network identified as the best one has difficulties in producing anatomical consistent images, even with the use of a proposed method to automatically crop the images in the abdominal ROI and of a Position-Based Selection method [147]. In fact, MR and CT images

exhibit important differences in texture which ease critic mechanisms of discrimination, due to the physical differences in acquisition. Conversely, for ceCT and CT domains, differences are more subtle, limited to certain anatomical parts and only in some 2D slices. In this context, the use of anatomically-paired images is a key element for discriminator specialization [121, 147].

In the abdominal region, there is a lack of spatial consistency: size and length of the organs, as well as their relative position, may vary a lot from one patient to another one. Simple matching strategies, such as the PBS method, without and with a step of 3D affine registration, fails in selecting good anatomically-paired images. Our proposed method with the use of Self-Supervised Body Regressor solves these problems, since it assigns the same label to slices exhibiting a similar anatomy. In addition, besides improving the selection of anatomically-paired slices, SSBR also allows us to proceed as in AC-GAN [104] forcing, together with the input addition and the binary mask, the anatomical consistency between input and synthesized output. The best qualitative and quantitative results (albeit the latter are limited to a few subjects) confirm the possibility of improving the structural consistency in unpaired ceCT-CT translation leveraging anatomical constraints.

Moreover, although the method is designed for the difficulties of ceCT-CT translation in the abdominal region, this is applicable on other translation tasks, such as MRI to CT or T1-w to T2-w, and other body sections. In the time available and given the goals of this thesis, we have not been able to conduct such experiments. Furthermore, given the possibility of our method to be used independently of the choice of generating network and discrimination mechanism, with a larger database and a more powerful GPU available, transformer-based methods [68] may be further explored. Again, the use of such methods with high demands on time and memory would be reasonable in case one aims to improve the contrast performance of synthetic ceCT images. Indeed training on non-pathological adults and using the trained network in inference on pathological children result in high fidelity contrast-free CT images in both contrast and anatomy, which is the CT modality usually not acquired. Figure 7.2 illustrates this, where we can see high anatomical coherence for all transformations but unrealistic contrast for synthetic ceCT images, in particular if a tumor is present.

The use of synthetic CT images with real ceCT images produced segmentation results on arteries and veins in line with the use of both real images, and is therefore considerable as a viable alternative to double X-ray exposure. Furthermore, using this technique on the entire database results in an even higher improvement in performance, in particular on the subjects and areas with higher heterogeneity. This technique has also proved useful for other structures with less contrast variability. The use of anatomically fidelity images for consistent match between images and reference segmentation is critical.

In my humble opinion, this combination of double CT modalities as an additional iconographic data augmentation is a useful technique to partially tackle the variability given by the contrast medium diffusion, and should be used to obtain the best performance in the segmentation of pediatric and pathological abdominal ceCT images.

Segmentation of renal tubular structures. The second most important group of structures in a complete renal 3D digital twin are the tubular structures: arteries, veins and ureters. Manual segmentation of such structures in ceCT images is the most time-consuming step in the model creation, and using basic image processing techniques such as thresholding and region growing (even in a locally adaptive manner) does not provide much benefit. In addition to the problem of contrast heterogeneity already addressed, this difficulty is also due to the

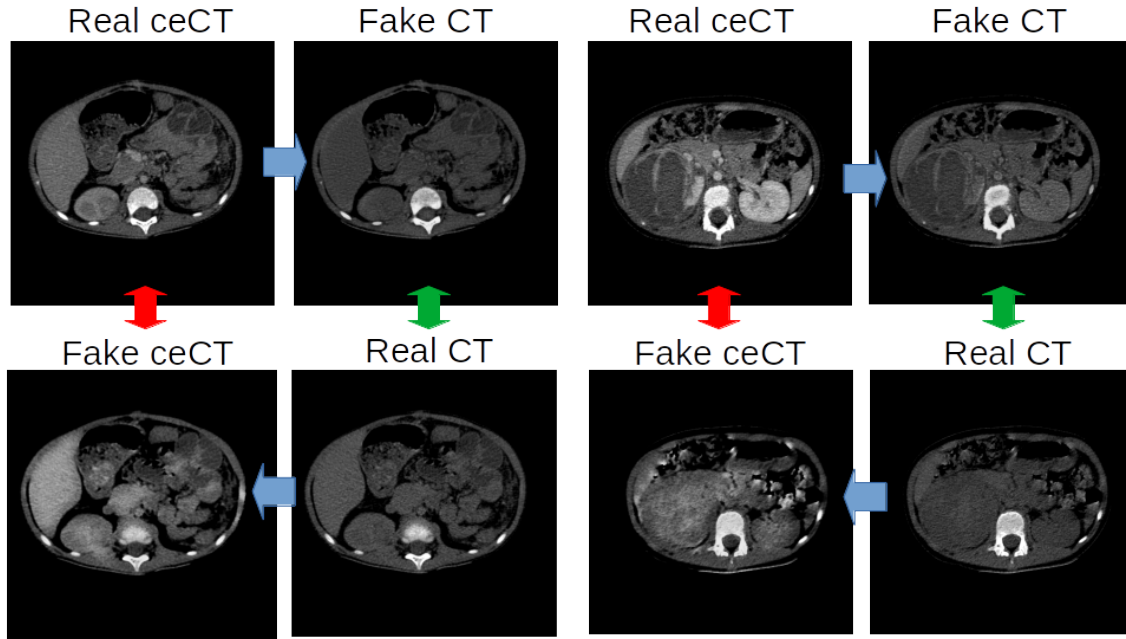


Figure 7.2: ceCT-CT translation obtained with CycleGAN trained on unpaired adults and used in inference on paired children, as explained in Section 4.2.3. Blue arrows: direction of transformation; green arrows : high anatomical and contrast fidelity; red arrows: high anatomical but low contrast fidelity.

arteriovenous phase acquisition in order to have all structures visible, including adjacent organs such as liver and spleen. This acquisition phase does not allow for an obviously different in contrast intensity both between arteries and veins and between blood vessels and renal parenchyma. An example is shown in Figure 7.3.

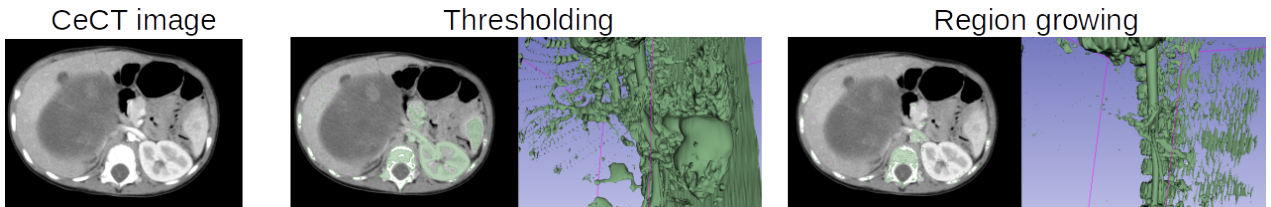


Figure 7.3: An example of ceCT image with arteriovenous injection where the use of thresholding (too many false positives) or region growing (difficulty in selecting starting points) does not provide much benefit compared to the manual segmentation.

Despite the lack of literature dedicated to renal tubular structures in pediatric patients, an extensive assessment of the methods used on adults allowed us to understand how the different issues related to such structures are addressed. However, none of the methods analyzed showed very high performance, except for specific acquisition CT modalities.

Due to the presence of few biphasic images we could not use a single dataset for both blood vessels and ureters. We therefore had to divide the tests into one database with early-phase acquisition images for segmentation of arteries and veins, and another with late-phase acquisition images for ureter segmentation (very limited in number). Since the late-phase acquisition only shows one well-contrasted and usually regular structure one might think that in these we could use the above-mentioned rule-based techniques. However, the new biphasic acquisition protocol (see Chapter 2) led us to develop a method suitable for an image with all

renal tubular structures with a similar contrast intensity.

The comparison of segmentation results of the state-of-the-art methods on the Necker PRAC database has allowed us to identify the best techniques among those assessed. Once again nnU-Net [61] turned out to be the most performing, and its simplicity of implementation allowed us to build a new method starting from this. The main limitation related to these methods appears to be related to the use of only voxel-wise loss functions. The information that these functions provide does not exploit the particular knowledge we have on the morphology of such tubular structures. A proposed solution in the literature [15, 134] is to weight the voxels according to their distance map. Personally, I find this solution useful only to cover intra-scale changes but not the strong inter-anatomy variation in pediatric patients, resulting in a better segmentation only of the contour voxels of larger structures such as aorta and cava vein. Furthermore, another problem found by analyzing these results is that of the interruption of the structures which makes manual correction even more tedious. All this is even more confirmed by the evaluation measures used, such as Dice score and Hausdorff distance, which are not really informative about the segmentation results.

The design of what we have called Tubular structures Loss function allows us to overcome these problems, ensuring a morphological similarity to the target tubular structure by leveraging the comparison between the eigenvalues of the Hessian matrix [89] and uninterrupted segmentation thanks to the use of Frangi’s vesselness [38]. The good results of this approach are due to three factors: (i) the use of Hessian eigenvalues method on the output image, i.e. the probability map; (ii) ordering such eigenvalues not by magnitude but via their associated normalized eigenvectors, matching those with smaller angles to each other; (iii) finally, the use of this loss function in a multi-scale approach via deep supervision. Furthermore, the Morphological similarity Loss function emerges as a useful quantitative measure. The qualitative results are those that most confirm the high results obtained with this method.

In order to further validate the proposed method, experiments on other databases and other tubular structures can be performed. Also, one idea is to use uroCT scanners (images very close to ceCT scanners with delayed phase) that can be gathered at Necker hospital, to expand the labeled ureters database in order to improve their segmentation performance and have a larger, and therefore reliable, test set. Furthermore a subdivision between calyces and true ureters should be performed. Although one might argue that then a semantic division between the branches of the blood vessels should also be made. However, I do not believe this would lead to improvements in segmentation results; on the contrary it would give the network further difficulties given the lack of absolute position due to the training by patches and not by the entire volume. If one would take this path of the semantic division of blood vessel branches, it would be interesting to analyze methods that combine CNN with Graph Neural Network [41, 129].

I found this task the most difficult to achieve and although the performances are significantly improved compared to state-of-the-art methods, there are still some subjects where the vascular tree is very difficult to automatically segment. However, in the same subjects the manual segmentation is even more complicated, and the results obtained still allow for a gain in time and in user interaction.

3D anatomical digital twin to help surgery. Considering the first goal (automatic segmentation of principal structures in a 3D renal model) achieved entirely for kidney and tumor and partially for tubular structures, I dedicated myself to making the proposed approaches

usable in a real clinical setting for the 3D digital twin creation from pediatric ceCT with renal tumors. Through the use of 3DSlicer [37] and Docker [98], I developed a user-friendly tool for clinicians that, in addition to segmentation methods, also leverages transfer learning methods from adults to segment bone, liver, and spleen (although the latter with low performance) and the use of STNs to find the bounding box of interest speeding-up inference time and improving performance.

Physicians, with whom I have collaborated, have found the simplicity sought in the use of such a plug-in, and early tests on time saved have shown promising results. A systematic quantitative assessment, such as the one in [21] on segmentation time and inter- and intra-subject segmentation variability, can be important to the scientific community. Such a study would allow for a better understanding of the benefits provided by the plug-in, in terms of (i) the total segmentation time over a larger database, (ii) the intra- and inter-subject segmentation variability and whether the use of automatic segmentation as starting point leads to more similar final segmentations, as we expect. Moreover, this study will allow us to know what remains to be done in order to solve the problem of automatic multi-structure renal segmentation in ceCT imaging.

The protocol for validating the 3D models through verification by both radiology and visceral surgery teams allows for high confidence in 3D models, and some cases have already confirmed the benefits brought by the models in both pre-operative planning and per-operative guidance. Also in order to further evaluate the advantages from a visualization and interaction point of view, a retrospective study is on-going.

The use of the 3D model also stands as a means of education for students or residents unaccustomed to projecting into their own mind what they will find in the operating room. Moreover it also emerges as a useful support for the preoperative discussion with the patient's family, allowing the physician to provide a clearer explanation of the procedure and what the complications might be.

From my point of view, the results achieved by this thesis are satisfactory given all the difficulties raised by pediatric and pathological ceCT images and stand as a more than good starting point. There is still work to be done on both the plug-in and the automatic segmentation in order to make this tool usable day-by-day by non-expert 3DSlicer users and thus to increase the number of patients who can benefit from the construction of the 3D digital twins. The hope is to increase more and more partial nephrectomies even in unilateral cases, saving the patient from the long-term risks associated with having only one functional kidney. In order to give even more value to the 3D digital twins, a prospective study needs to be performed, involving multiple centers and a large number of patients (at least 100 patients).

7.2 Perspectives

The work presented in this PhD thesis also paves the way for several open issues and long-term questions for the scientific community. Such perspectives for each of the contributions are here discussed.

Segmentation of kidneys and renal tumors. To further improve both direct inference and transfer learning from adults weights to children image, one can exploit a geometric domain adaptation with both Spatial Transformer Networks and non-linear transformations. As we have seen, in fact, while the kidneys are larger in adults than in children, the tumors have the

opposite behavior. A proposition can be a double deformation modeled as a composition of a linear and a non-linear deformation that adapts a segmentation network trained on a large adult dataset to a small pediatric dataset, as shown in Figure 7.4. It is based on the idea of CycleGAN [157] but instead of intensity (i.e. iconographic) transformations, we propose to use geometric deformations. It is important to note that in this case both datasets are needed, so other domain adaptation methods should be tested, such as the domain adversarial training [40]. From a mathematical point of view, we propose to model the differences between samples in X (pediatric images) and in Y (adult images) by using a composition of two deformations: a linear deformation ϕ^A and a non-linear one ϕ^N : $\phi = \phi^N \circ \phi^A$, so that $\phi_1(x_i) \in Y$ and $\phi_2(y_j) \in X$, where x_i denotes a pediatric image (in X) and y_j an adult image (in Y), and with index 1 used for transformations from X to Y , and index 2 for transformations from Y to X .

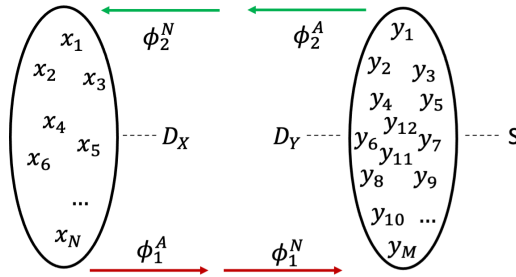


Figure 7.4: Pediatric x_i and adult y_j images are transformed into the other domain using a composition of a linear ϕ_1^A (resp. ϕ_2^A) and non-linear ϕ_1^N (resp. ϕ_2^N) deformations. A segmentation network S , previously trained on the adult data set, is used to optimize the two deformations. Two adversarial discriminators, D_X and D_Y , are also trained as in Cycle-GAN [157] to drive the resulting transformed images to be indistinguishable from the target domains.

The first linear deformation ϕ^A should take into account global differences in terms of size and pose between pediatric and adult images. The second non-linear transformation should mainly account for local differences, in particular for the relative size and shape differences between healthy structures and tumor. Linear deformations could be modeled using STN [65] and the non-linear deformations using VoxelMorph [4] or diffeomorphic autoencoders [9]. Both linear and non-linear deformations should be optimized in order to correctly transfer the segmentation network from the adult domain to the pediatric one. In order to do that, we could update the parameters of ϕ_1 by minimizing $d(\phi_1(G_{x_i}), S(\phi_1(x_i)))$, where G_{x_i} is the reference segmentation for x_i , S is the segmentation produced by the network. However, this strategy has two flaws. First, one would need to interpolate the reference segmentation, which should not be touched by definition. Secondly, at inference time, one would like to have the segmentation result in the original pediatric space (i.e. X) and not in the adult domain (i.e. Y). To overcome these problems, we propose different solutions. The first one is similar to what we proposed in our ISBI paper [P3] presented in Section 3.4. Here the U-Net is trained by transforming the obtained segmentation in an inverse way $\phi_1^{-1}(S(\phi_1(x_i)))$ in order to compare it with G_{x_i} . Having already verified the possibility of easily inverting computationally ϕ_1^A , we still have to find possible non-linear transformations that are always invertible and computationally feasible for ϕ_1^N . We did some preliminary tests using Thin-Plate Spline (TPS) interpolation [10] shown in Figure 7.5. Satisfactory results using the renal ROI are obtained only when the tumors have already a similar size, while a pre-segmentation of kidneys and renal tumors and more parameters are needed if the pediatric tumor is considerably larger than

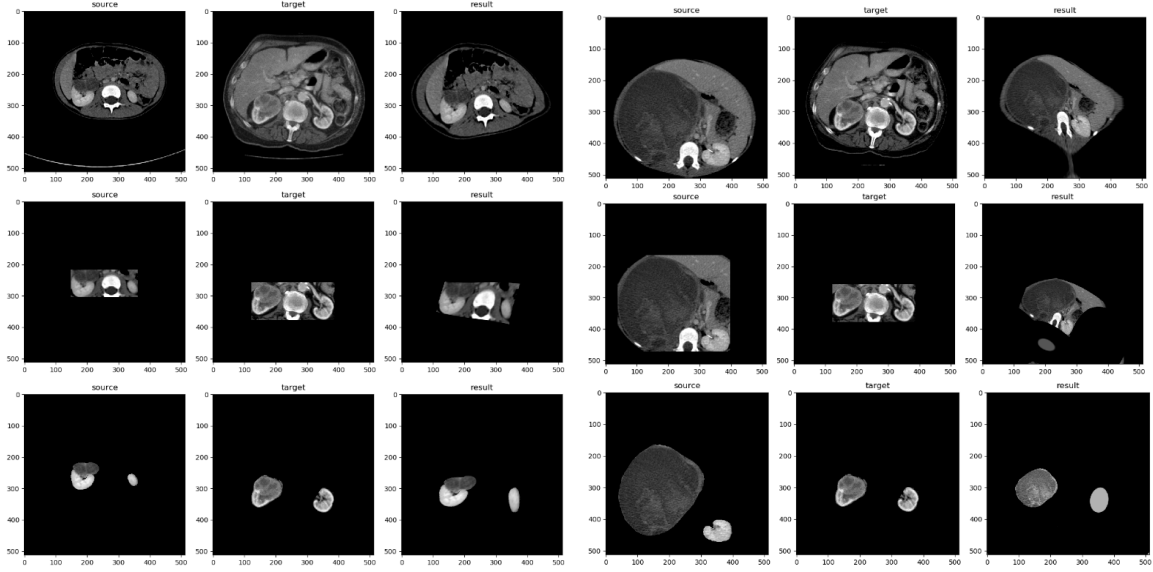


Figure 7.5: Preliminary results using affine linear transformation and TPS [10] non-linear transformation with 4 (three left columns) and 8 (three right columns) parameters. Top: entire abdomen, middle: renal ROI, bottom: pre-segmented kidneys and renal tumor.

adult one. Moreover TPS is not inverse-consistent [10] and one should implement a modified algorithm that uses an approximating spline. This would allow us to use our proposition, and there would be no need to have the adult database as well (but only their pre-trained weights). Another possible solution, which would instead require also the adult database, is illustrated in Figure 7.4, and consists in leveraging the cycle consistency idea from CycleGAN by estimating ϕ_1, ϕ_2 minimizing $\sum_{i=1}^N d(G_{x_i}, \phi_2(S(\phi_1(x_i))))$ where N is the number of images of domain X . In this way, ϕ_1 would be optimized to correctly bring x_i in the Y domain and ϕ_2 would be optimized to correctly bring the segmentation in the original pediatric domain X . In this case, ϕ_2 must be very performing to obtain reliable results. A third solution would be the one presented in [145] with the transformed images used as data augmentation in two separate segmentation networks in order to improve the performance of both networks.

Tumors in adults and children often evolve differently, regardless of the body region. Therefore, the method proposed here is easily applicable to other scenarios than renal tumors, as well as to other acquisition modalities. It could also be used for efficient transfer learning between patients of the same domain (i.e. adults or children) which have tumors that by evolution or stage of discovery have different sizes.

Cross-domain CT image translation using CycleGAN with anatomical constraints.

Given the satisfactory results for synthetic ceCT images in non-pathological adult subjects, the application of these generated images in a real clinical setting for simulated injection and its clinical pro and cons should be evaluated. Researchers seem to point more and more towards a virtual contrast medium injection. Feasible studies for brain MR images were done by some authors [76, 14], and they led to conflicting results. In addition to having results still far from clinical use - but already useful for decreasing the dose of contrast injected into the patient - it is clear that there are really many particular cases in which this technique cannot be used.

We propose to proceed in a similar way for synthetic abdominal ceCT images but we

want to answer more questions which do not seem to have been addressed. First, according to medical doctors who are part of the project team, a real ceCT is essential for clinical diagnosis, for several reasons. An example is that their decisions are based on the symmetry of the contrast medium diffusion to detect the presence of capsules on tumors or kidneys, which is given by the physiology of such structures and cannot be virtually reproduced. So we have to determine if there are potential clinical applications not linked to the direct use for diagnosis. A first example could be a patient who undergoes a contrast-free CT for a particular exam that does not need the contrast enhancement. In this case, we can produce a synthetic ceCT which may suggest to actually acquire a real ceCT, according to certain features to be assessed via a joint study with physicians.

To sum up, our idea is to show the limits of virtual injection in medical images but also how far we can go, what situations could be simulated. A large-scale discussion with medical experts on this topic is important.

Segmentation of renal tubular structures. To date, I find it really difficult to improve even more the segmentation of renal tubular structures in pediatric and pathological arteriovenous ceCT via deep learning techniques. An idea could therefore be to combine these techniques with rule-based methods as a post-processing step, refining the automatic segmentation obtained by the CNN. In order to leave the choice on the application of this method to the user, we propose to create a new Segment Editor effect for the 3DSlicer [37] plug-in (presented in Section 6.1) dedicated to tubular structures. Our idea is the development of an Automatic Locally Adaptive Region Growing [54], that we call “A-LARG”. Once the segmented structure is selected on the Segment Editor (e.g. arteries), the algorithm could proceed as follows:

1. a skeletonization is applied to the segmentation (Figure 7.6a);
2. the end and bifurcation points are extracted and selected as initial points for the region growing (Figure 7.6b);
3. the previous points are used to transform the skeleton to a graph (Figure 7.6c);
4. for each branch of the graph the standard deviation of contrast intensity of the corresponding voxels on the input ceCT image is calculated (Figure 7.6d);
5. for each initial point a Locally Adaptive Region Growing [54] is performed using as contrast tolerance the maximum of the standard deviation among the limbs branching off from that point.

This algorithm can take advantage of the segmentation obtained from the proposed method for tubular structure segmentation that forces the segmentation to be uninterrupted. One problem could be caused by the contrast heterogeneity in ceCT images which induces outliers in the contrast tolerance calculation, as some preliminary testing has already shown. Moreover, a stop condition should also be implemented.

In addition to possibly improving the segmentation of arteries, veins and ureters, this method would allow physicians to segment other vessels contiguous to those, such as the mesenteric artery, or hepatic arteries and veins, in case they are needed for a better surgical planning and guidance.

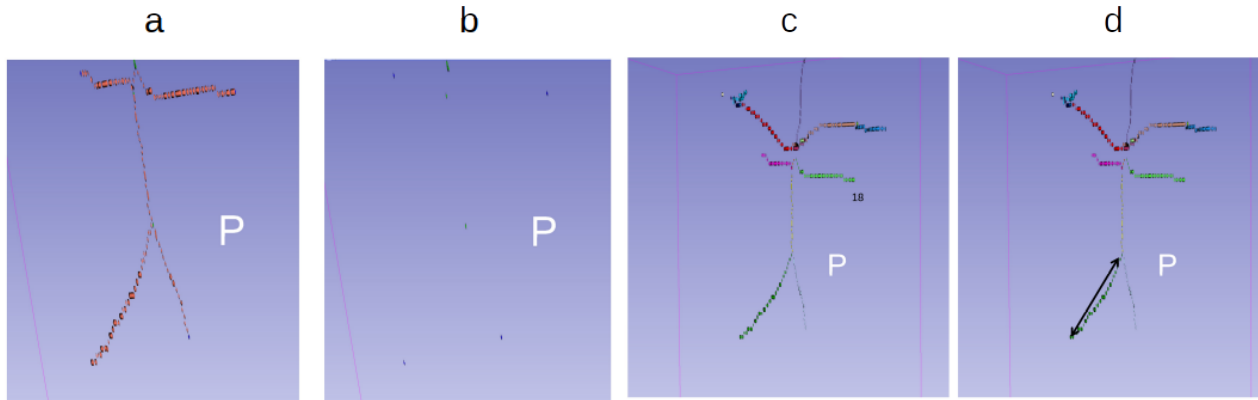


Figure 7.6: Steps of A-LARG algorithm. a: skeleton of initial tubular structure segmentation; b: end (blue) and bifurcation (green) points; c: from skeleton to graph, each branch has a different colors; d: single branch highlighted by the arrows.

The proposed A-LARG algorithm can be applied to any tubular structure that is already fully or partially segmented, either manually or automatically using deep learning techniques (as in our case). Other applications may be pelvic blood vessels, nerves or lung bronchial tree.

3D anatomical digital twin to help surgery. To further complement the information provided by the 3D model, and to give it even more power than using 2D imaging alone, an interesting approach is to combine information acquired from multiple modalities. As shown in Section 6.2, a first example is the use of both ceCT and MR, where the former provides more detailed information about the vessels and ureters (with a biphasic acquisition by combining arteriovenous and excretory phases), while the second gives more detailed information about renal parenchyma (distinguishing internal and external) and tumors. Another example is combining anatomical information with functional information. If the former is extracted from structural images (MR or ceCT), the latter can be extracted from ultrasonographic (US) images. The idea would be to train the network on pre-operative US images, which are easier to register following specific protocols, and then perform automatic real-time segmentation on US images acquired during surgery and inject this information into the pre-computed 3D anatomical model from ceCTs.

The use of US also opens the direction of longitudinal follow-up of tumor evolution before, during, and after the 4-week chemotherapy (as recommended by the Umbrella SIOP protocol [12]) via an automatic segmentation algorithm.

Moreover, another direction already taken by the IMAG2 team at Necker hospital, is the automatic super-imposition of the 3D anatomical model on the images acquired by the operating camera. This is particularly important for laparoscopy performed via the “da Vinci Surgical System, Intuitive®”, due to the even restricted field of view and feeling of immersion given by the system. In fact, the surgeon, thanks to the reference points given by the 3D model, would find a greater feeling of comfort and safety in the surgical operation, as already demonstrated in [109, 133], being also able to identify structures or locate organs that are not yet visible. A part of the IMAG2 team is working on the automatic segmentation and subsequent creation of the 3D point cloud model from stereo images acquired with the robot laparoscope, i.e. stereo camera (Figure 7.7 from a to c). Then, the idea is to perform an alignment with the 3D point cloud and the pre-operative 3D anatomical digital twin using

sub-cloud of points corresponding to the organs to be aligned (Figure 7.7 d and e). To this end, the position of the camera should be deducted in order to be able to reproject the 3D model in the operating field of view of the camera. These steps are summarized in Figure 7.7.

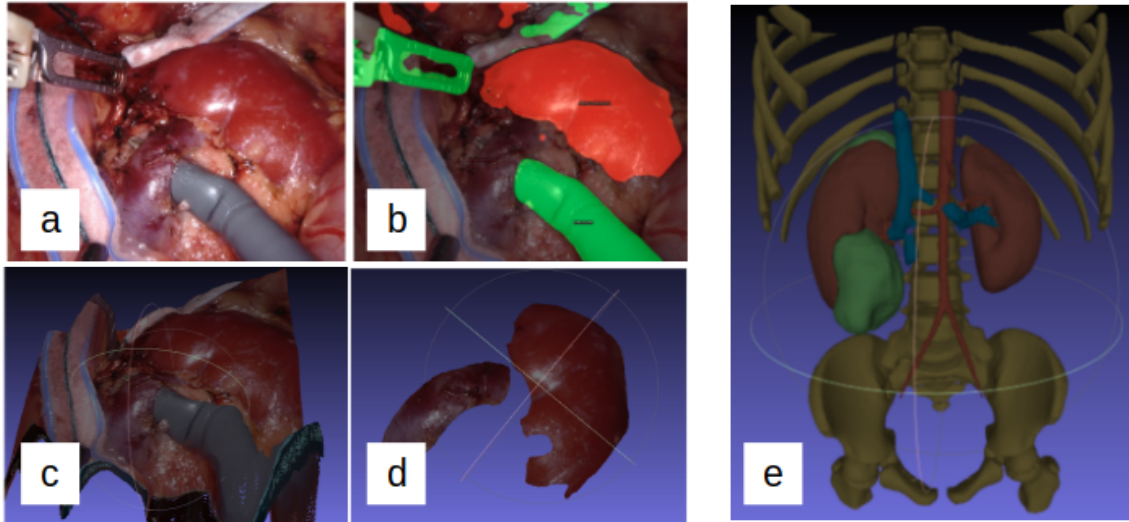


Figure 7.7: Steps for the super-imposition of the 3D anatomical model on the images purchased by the operating camera. a: laparoscopic image; b: segmentation of laparoscopic image; c: 3D point cloud of b; d: sub-cloud of target organ; e: pre-operative 3D digital twin.

The final goal of this idea is to provide surgeons with an augmented reality tool that allows them to know exactly where they are in the patient's body by registering the complete pre-operative 3D model with the per-operative laparoscopic image.

The use of 3D models both in their digital version addressed in this thesis and in their physical version, through the use of 3D printers, offer other numerous application insights. 3D impressions are increasingly used in maxillo-facial and orthopedic surgery [30] for the production of cutting guides that reduce operating time and increase surgical precision. The 3D digital twin, in addition to being less expensive, turns out to be more useful for surgical applications such as those shown in this thesis, where vessels or excretory pathways enter inside the organs being examined. Its use in adult abdominal-visceral surgery is growing, but automated tools for segmentation such as the one we presented are still limited [94, 112]. If we focus on the pediatric scenario this lack is even stronger, because the children population requires special attention due to the specificity of the pathologies concerned and the importance of precision and minimal-invasive surgery, whose results will have a very long-term impact. There are few studies that have evaluated the relevance of these tools in pediatric surgery and they most often concern surgery for renal tumors [45, 106] or modeling of congenital heart disease [101]. Other applications in fields such as oto-rhino-laryngeal or pelvic surgery should be considered. Moreover, the use of tractography techniques for modeling neural traits from diffusion-weighted imaging (DWI) could be also merged with the 3D anatomical model, improving pre- and per-operative neuroblastoma surgery compared to the conventional 2D images, both in the relationships between the tumor (around the kidney in this case) and the renal tubular structures, and in the relationships between the tumor and the nerves. Furthermore, this method could also be used for studies on adults, for example affected by endometriosis.

Publications

Preliminary to the thesis

- P1. [OHBM-19] Giammarco La Barbera, Isabelle Bloch, Gonzalo Barraza, Catherine Adamsbaum, and Pietro Gori, “Robust segmentation of corpus callosum in multiscanner pediatric T1-w MRI using transfer learning,” in OHBM 2019 - Organization for Human Brain Mapping, Jun 2019, Rome, Italy. [⟨hal-02934231⟩](#)

Out-of-thesis’ subject

- P2. [MICCAI-20] Minhao Hu, Matthis Maillard, Ya Zhang, Tommaso Ciceri, Giammarco La Barbera, Isabelle Bloch and Pietro Gori, “Knowledge distillation from multi-modal to mono-modal segmentation networks,” MICCAI 2020 - Medical Image Computing and Computer Assisted Intervetion, Oct 2020, Lima, Peru. pp.772-781. [⟨hal-02899529⟩](#)

International conferences

- P3. [ISBI-21] Giammarco La Barbera, Pietro Gori, Haithem Boussaid, Bruno Belucci, Alessandro Delmonte, Jeanne Goulin, Sabine Sarnacki, Laurence Rouet and Isabelle Bloch, “Automatic size and pose homogenization with Spatial Transformer Network to improve and accelerate pediatric segmentation,” in ISBI 2021 - IEEE International Symposium on Biomedical Imaging, Apr 2021, Nice, France. pp. 1773-1776. [⟨hal-03131980⟩](#)
- P4. [BMVC-22] Giammarco La Barbera, Haithem Boussaid, Francesco Maso, Sabine Sarnacki, Laurence Rouet, Pietro Gori and Isabelle Bloch “Anatomically constrained CT image translation for heterogeneous blood vessel segmentation,” in BMVC 2022 - British Machine Vision Conference, to be held in Nov 2022, London, United Kingdom. [⟨hal-03797472⟩](#)

French conferences

- P5. [SFCP-21] Rani Kassir, Giammarco La Barbera, Alessandro Delmonte, Cécile Lozach, Thomas Blanc, Isabelle Bloch and Sarnacki Sarnacki. “Importance d’une modelisation 3D multimodale des tumeurs renales de l’enfant: a propos d’un cas de nephroblastome bilatéral,” in SFCP 2021 - Congrès annuel de la Société Française de Chirurgie Pédiatrique, Oct 2021, Online.
- P6. [GRETSI-22] Giammarco La Barbera, Haithem Boussaid, Francesco Maso, Sabine Sarnacki, Laurence Rouet, Pietro Gori and Isabelle Bloch, “Synthèse non supervisée d’images ceCT-CT sous contrainte anatomique,” in GRETSI 2022 - Colloque Francophone de Traitement du Signal et des Images, Sep 2022, Nancy, France.

Journal articles in preparation

P7. Based on the content of Chapter 5, to be submitted to Medical Image Analysis.

P8. Based on the content of Chapter 6, to be submitted to Medical Pediatric Radiology.

Appendix A

Evaluation measures

For the quantitative evaluation of the segmentation results, we compute two different categories of measures, as defined in [127]:

Spatial overlap based measures. All measures from this category can be derived from the four basic cardinalities of the confusion matrix, namely the true positives (TP), the false positives (FP), the true negatives (TN), and the false negatives (FN). These values can be expressed for a single class c (i.e. structure) as follows:

$$\begin{aligned} TP &= \sum_{j=1}^M |\hat{P}_j \cdot R_j| & TN &= \sum_{j=1}^M |(1 - \hat{P}_j) \cdot (1 - \hat{R}_j)| \\ FP &= \sum_{j=1}^M |\hat{P}_j \cdot (1 - R_j)| & FN &= \sum_{j=1}^M |(1 - \hat{P}_j) \cdot R_j| \end{aligned} \quad (\text{A.1})$$

where M is the number of voxels in the image, R_j is the value of the voxel j of the one-hot encoded reference segmentation R (same used during training), and is a value in $\{0, 1\}$, and \hat{P}_j is the value (in $\{0, 1\}$) of the voxel j of the one-hot encoded predicted segmentation (we first apply an *argmax* operation on the classes C of probability map P_{cj}). From this category we used the Dice score (Equation A.2) as the most used measure to validate medical volume segmentations [127], Precision (Equation A.3) and Recall (Equation A.4), useful for understanding the amount of manual corrections to be made (how much to delete and how much to add, respectively). These scores are defined in percentage as:

$$Dice\ Score[\%] = \frac{2TP}{2TP + FP + FN} \cdot 100 \quad (\text{A.2})$$

$$Precision[\%] = \frac{TP}{TP + FP} \cdot 100 \quad (\text{A.3})$$

$$Recall[\%] = \frac{TP}{TP + FN} \cdot 100 \quad (\text{A.4})$$

Spatial distances. These measures take into consideration the spatial position of voxels and they are recommended when the boundary delineation (contour) of the segmentation is of importance [127]. Here we used the 95th percentile of Hausdorff distance (95HD), that is slightly more stable to small outliers than the standard Hausdorff distance, which is an

indicator of the largest segmentation error and is commonly used in biomedical segmentation challenges [127]. The distance unit is the same as for the spacing of elements along each dimension, which is usually given in *mm*. For two point sets, i.e. voxels in a space coordinate system, R and P , the HD is defined as:

$$HD = \sup \left\{ \sup_{p \in P} \inf_{r \in R} \|p - r\|_2, \sup_{r \in R} \inf_{p \in P} \|p - r\|_2 \right\} \quad (\text{A.5})$$

Appendix B

Details on the parameters used in data augmentation and affine registration

Table B.1: Data Augmentation parameters in nnU-Net [61]. *Zero centered additive Gaussian noise is added to each voxel in the sample independently. **Images are downsampled by a low-resolution factor using nearest neighbor interpolation and then sampled back up to their original size with cubic interpolation.

Spatial data augmentation	Application probability	range of values
Scaling	0.2	[0.7,1.4]
Rotation	0.2	[−30,30]
Mirror	0.5 along all axes	-
Iconographic data augmentation	Application probability	range of values
Contrast	0.15	[0.7,1.5]
Brightness	0.15	[0.7,1.3]
Gamma	0.15	[0.7,1.5]
Gaussian Noise*	0.15	[0,0.1]
Gaussian Blur	0.2	[0.5,1.5]
Low Resolution simulation**	0.25	[1,2]

Table B.2: Some of the parameters for 3D affine registration using SimpleITK-SimpleElastix [96].

Parameter name	Parameter value
Final BSpline Interpolation Order	2
Interpolator	Linear Interpolator
Maximum Number Of Iterations	32
Maximum Number Of Sampling Attempts	8
Metric	Advanced Mattes Mutual Information
Number Of Samples For Exact Gradient	4096
Number Of Spatial Samples	4096
Optimizer	Adaptive Stochastic Gradient Descent
Registration	Multi Resolution Registration
Resample Interpolator	Final BSpline Interpolator

Appendix C

Supplementary material for kidney and renal tumor segmentation

Transfer learning from adults to children

Table C.1: Quantitative results (mean and standard deviation of Dice score) of transfer learning on children using weights of 3D nnU-Net trained on KiTS database [53]. In italics in the first column are the same results as the ones shown in Table 3.3. Increasingly satisfactory results are obtained as the number of fine-tuned blocks increases.

Blocks re-trained	DS[%] Kidney	DS[%] Tumor
<i>Direct inference (weight frozen)</i>	20.83 (35.55)	18.29 (35.73)
<i>First 2 and last 2</i>	53.38 (25.84)	51.05 (31.76)
Bridge and last	57.39 (29.59)	59.20 (22.42)
First, bridge and last	49.69 (36.08)	40.66 (31.59)
First 2, bridge and last	51.05 (31.84)	53.38 (25.76)
Bridge, first of decoder, last	74.59 (7.31)	58.21 (25.50)
First 2 blocks	72.28 (17.94)	64.76 (26.67)
All decoder	76.90 (11.38)	75.33 (21.92)
<i>Bridge and all decoder</i>	81.75 (7.18)	75.79 (23.24)
<i>Entire network</i>	84.99 (6.38)	81.08 (23.01)

STN to homogenize pose and size

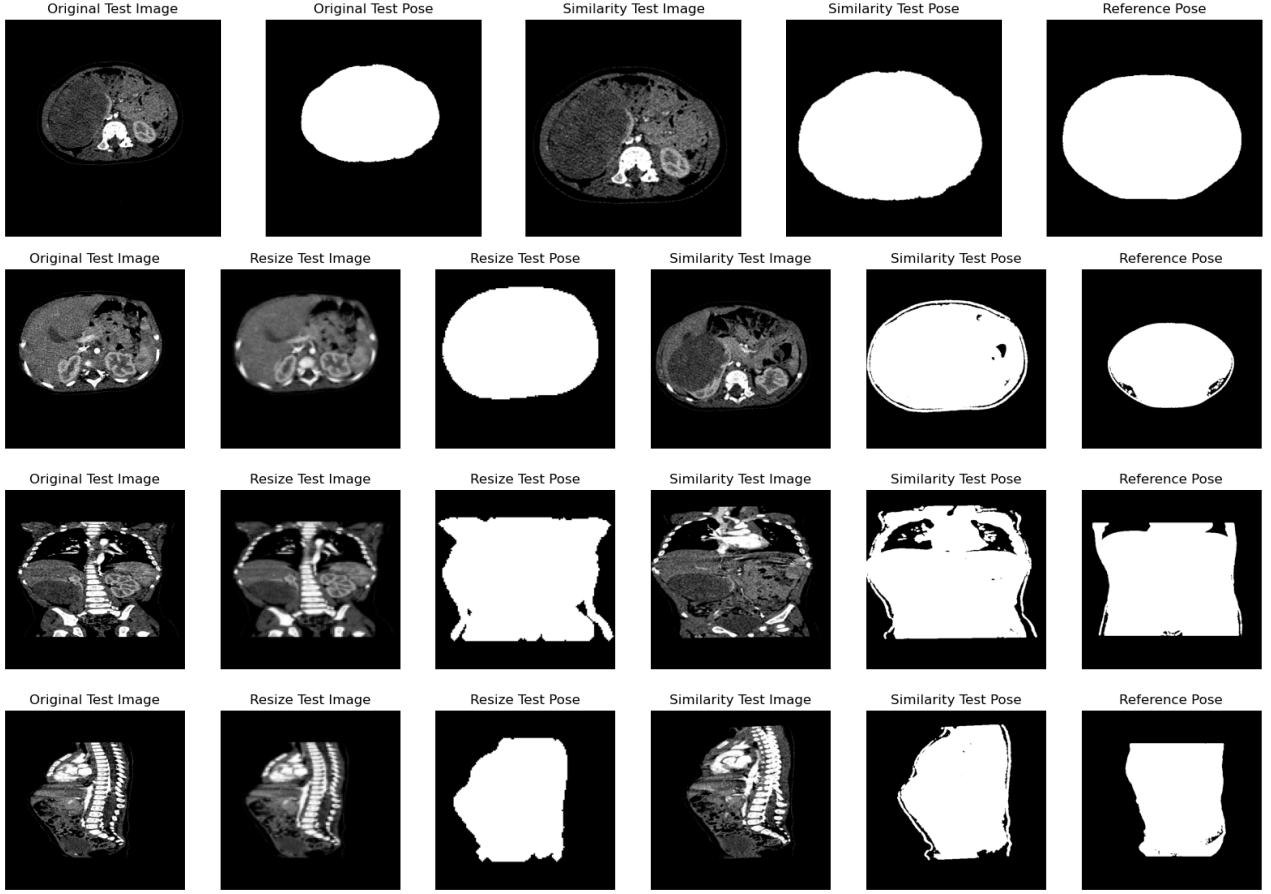


Figure C.1: Examples of results of STN^1 for pose and size homogenization in 2D (first row) and in 3D (last three rows in axial, coronal and sagittal views). It can be noticed how the homogenization in 2D (first row) is more performing than in 3D (second row) taking into account the same axis; moreover the latter seems to homogenize only the pose and not the size. It is also confirmed how resampling from $512 \times 512 \times 512$ to $128 \times 128 \times 128$ does not lead to a loss of detail that is significant for the task.

STN for ROI cropping

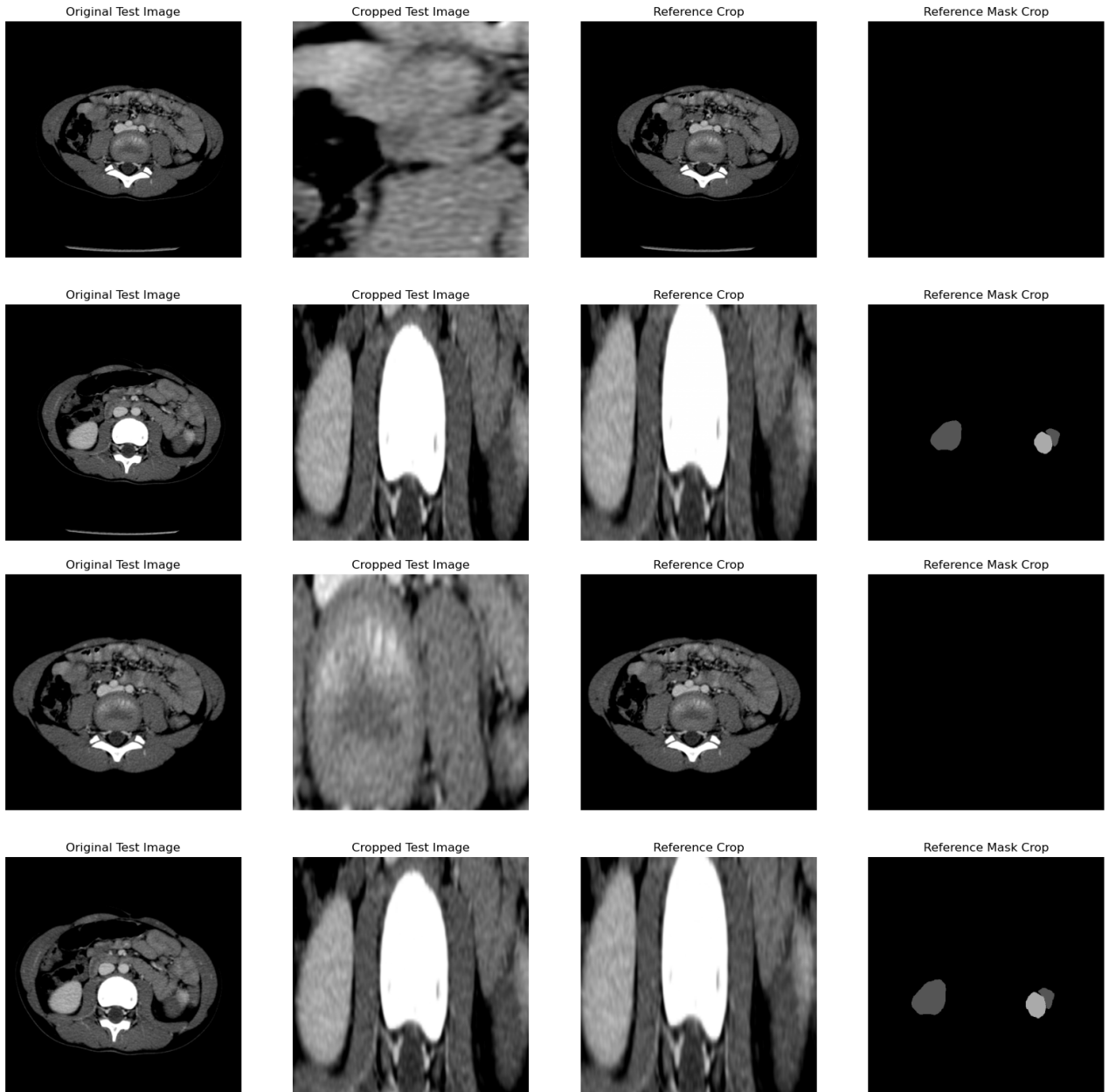


Figure C.2: Examples of results of STN^2 for 2D cropping using FasterR-CNN [114] as backbone. Without (first two rows) and with (last two rows) the use of STN^1 as pre-step. The results improve slightly with the use of STN^1 as the first step but the network has always a tendency to detect bounding boxes even if no target structure is present in the images.

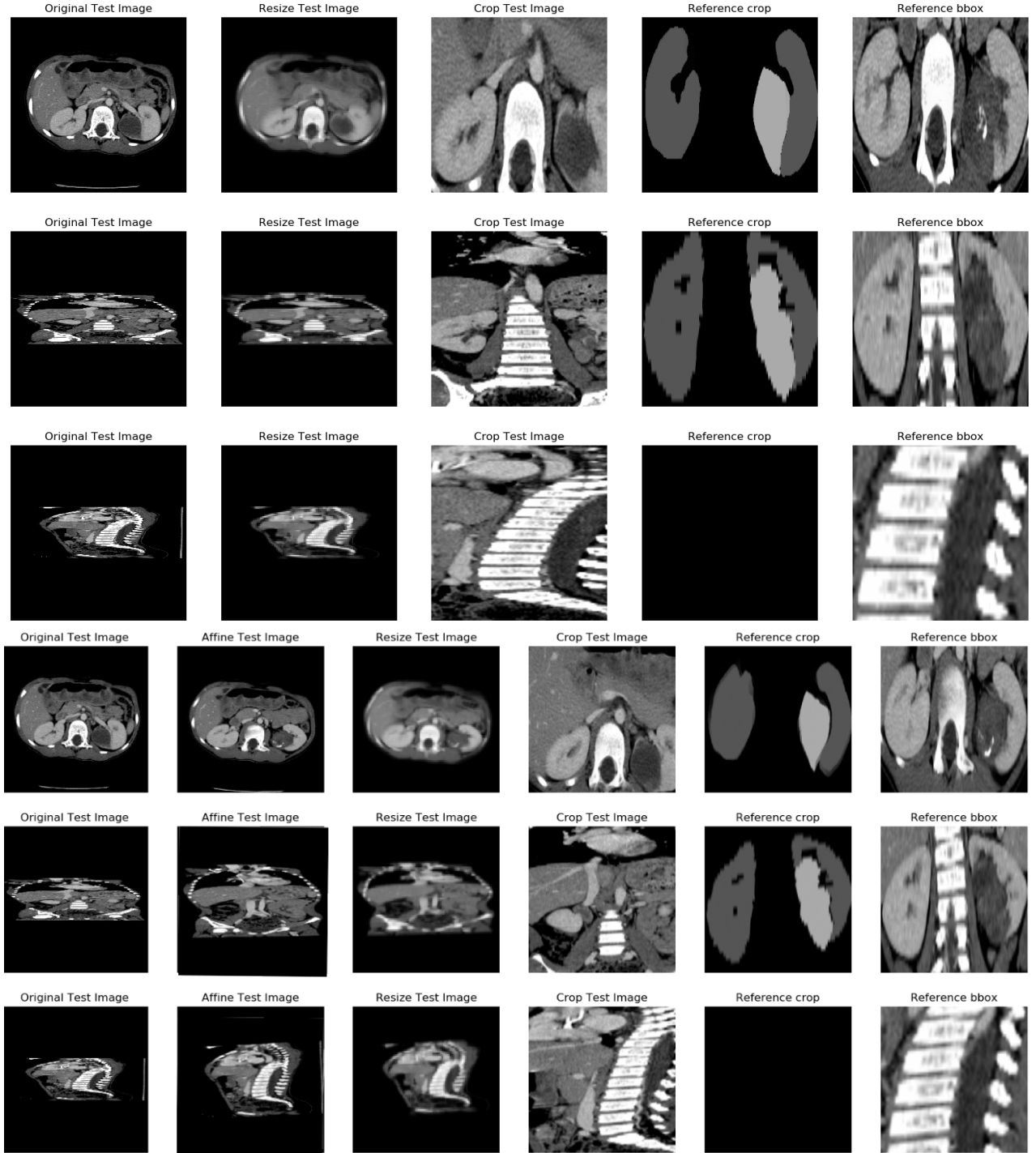


Figure C.3: Examples of results of STN^2 for 3D cropping using original STN [65] as backbone. Without (first three rows in axial, coronal and sagittal views) and with (last three rows) the use of STN^1 as pre-step. The examples show the non-satisfactory results of using this technique in 3D because of the limited number of patients to properly train the network. Using STN^1 as a first step does not lead to significant improvements.

Appendix D

Quantitative results on state-of-the-art methods for ceCT-CT translation

For an additional objective evaluation of generator performance on unpaired datasets, the use of the Fréchet Inception Distance (FID) [57] is now very common. Real and synthesized images are fed to the same network, usually the InceptionV3 model pre-trained on ImageNet. Then, the distributions of the features of the two sets of images are compared. One usually uses the features of one of the deepest layers. The idea is to use the first two moments, mean and covariance matrix, to compare the two distributions, and the FID is defined as:

$$\text{FID} = \|\mu_{real} - \mu_{fake}\|^2 + \text{Tr}(\Sigma_{real} + \Sigma_{fake} - 2(\Sigma_{real}\Sigma_{fake})^{1/2}), \quad (\text{D.1})$$

where μ_{fake} and μ_{real} are the mean of the generated fake and real images, Σ_{fake} and Σ_{real} are the relative covariance matrices, and Tr denotes the trace of a square matrix. A low FID should indicate that the two distributions are similar, namely that generated and real images should come from a similar distribution. However, patterns and representations learned from ImageNet may not be helpful in identifying useful and discriminative representations in medical images. Furthermore, FID only compares the first two moments of the distributions, which may be misleading or not informative enough in some cases (for instance if the distributions are not Gaussian). For these reasons, we do not consider these measures to be completely reliable, and they are not included in Chapter 4.

To overcome the previous issues, we propose two new measures for quantitative assessment of unpaired image-to-image translation. Let the *residual map* RM be the difference between the output and the input images of the generator, normalized between 0 and 1, where 1 means the maximum addition (or removal) of contrast. We then call *action region* RM_t the binary mask created by thresholding RM at 0.5. Let M be a manually segmented mask showing the area that should have the greatest variation of contrast (prior anatomical knowledge). We then define two quantitative measures. The “recall of action region”, which is a measure of completeness, showing how much of the target region has been changed, is defined as $R = \frac{|M \cap RM_t|}{|RM_t|}$, where $|\cdot|$ means number of pixels. The second measure, that we call “precision of the residual map”, measures if most of the changes in the output image have been concentrated in the correct parts of the image. It is defined as: $P = \frac{\sum_{x \in M \cap RM_t} RM(x)}{\sum_{x \in RM_t} RM(x)}$, where $RM(x)$ refers to the value of the pixel x in RM . The limitation on this method lies in the need of manual segmentations, which are really hard to correctly perform in CT images. Some examples are shown in Figure D.1.

Table D.1 provides the quantitative results (mean and standard deviation) for the some of the state-of-the-art experiments shown in Figure 4.6, that confirmed what was inferred from the qualitative results. However, we used only 10 2D images for each domain to calculate R and P, while 360 2D images for each domain were used for the computation of FID.

Table D.1: Results (mean and standard deviation) of R and P (see text for details) on 10 images for each domain, and of FID [57] on 360 images for each domain. The use of PatchGAN as discriminator mechanism is the one that allows to generate synthetic images with a distribution more similar to the real ones, while the use of Res-Net as generator network allows for more complete and accurate contrast adjustments. However, without anatomically constraining the synthetization, all methods tend to modify many parts of the image that should not be modified.

Quantitative Measure	UNIT <i>G</i> : U-Net <i>D</i> : PatchGAN	UNIT <i>G</i> : U-Net <i>D</i> : Wass. Loss	CycleGAN <i>G</i> : U-Net <i>D</i> : PatchGAN	CycleGAN <i>G</i> : Res-Net <i>D</i> : U-Net	CycleGAN <i>G</i> : Res-Net <i>D</i> : PatchGAN
ceCT2CT Recall (\uparrow)	0.81 (0.18)	0.87 (0.14)	0.81 (0.36)	0.59 (0.35)	0.85 (0.17)
ceCT2CT Precision (\uparrow)	0.15 (0.04)	0.16 (0.04)	0.14 (0.11)	0.22 (0.11)	0.15 (0.05)
FID CT domain (\downarrow)	219.53	238.07	153.2	349.54	118.39
CT2ceCT Recall (\uparrow)	0.77 (0.07)	0.71 (0.08)	0.68 (0.11)	0.81 (0.06)	0.71 (0.22)
CT2ceCT Precision (\uparrow)	0.09 (0.04)	0.09 (0.05)	0.10 (0.08)	0.02 (0.01)	0.10 (0.06)
FID ceCT domain (\downarrow)	203.1	257.67	180.33	270.79	160.65

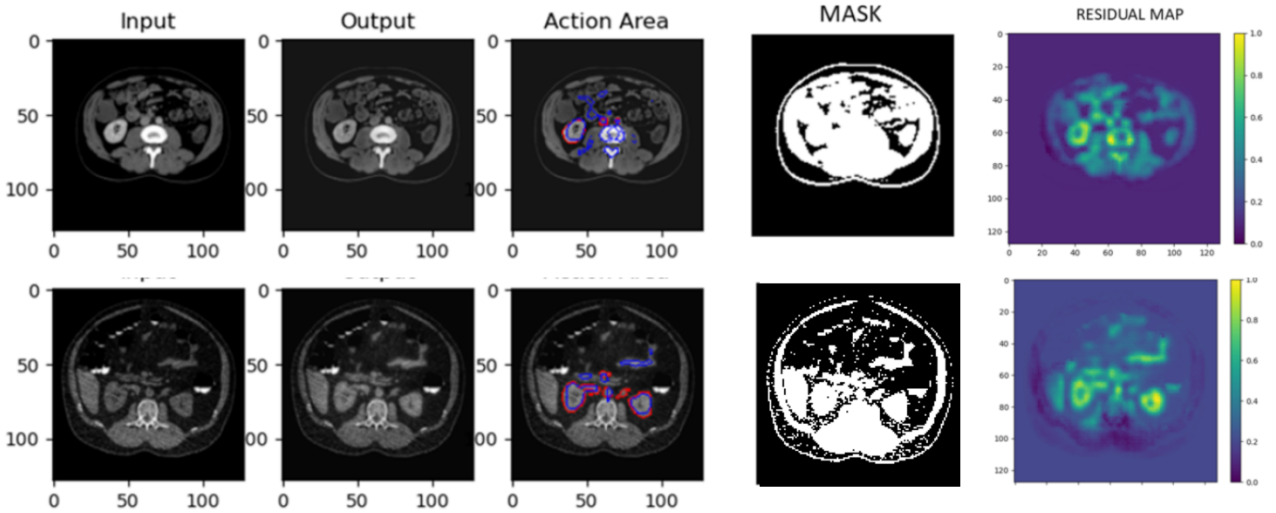


Figure D.1: Residual map and action area (detailed in the text). First two rows: from ceCT to CT. Second two rows: from CT to ceCT. In the images showing action area, the manual segmentation is displayed in red, and the automatic one in blue. In these two examples we can see that the action area coincides for most of the manual segmentation but also acts on many other areas.

Appendix E

Supplementary material for segmentation of tubular structures

Differentiability of Morphological similarity Loss

According to the first Magnus' theorem [91]:

“Let H_o be a real symmetric $n \times n$ matrix. Let W_o be a normalized eigenvector associated with a simple eigenvalue λ_o of H_o . Then a real-valued function λ and a vector function W are defined for all H in some neighborhood $N(H_o) \subset \mathbb{R}^{n \times n}$ of H_o , such that: $\lambda(H_o) = \lambda_o$, $W(H_o) = W_o$, and $HW = \lambda W$, $W'W = 1$, $H \in N(H_o)$, where W' denote the transpose of W . Moreover, the functions λ and W are ∞ times differentiable on $N(H_o)$, and the differentials at H_o are: $d\lambda = W'_o(dH)W_o$ and $dW = (\lambda_o I_n - W_o)_+^+(dH)W_o$. Equivalently, the derivative at H_o for λ is given by: $\frac{\partial \lambda}{\partial (\text{vec} H)'} = W'_o \otimes W'_o$ or $\frac{\partial \lambda}{\partial H} = W_o W'_o$, where $\text{vec} H$ denotes the column vector that stacks the columns of H one underneath the other, and \otimes denotes the Kronecker product.”

This means that if the Hessian matrix H is real symmetric, as in our case, the eigenvalues function λ is differentiable in the neighbor of H via the product of the normalized eigenvectors W_o (associated with the simple eigenvalue λ_o of H) with its transpose.

To give further details, the derivative of the cost function $MsLoss$ for a single predicted voxel P_m with respect to a parameter Z_p of the network Z can be written as:

$$\frac{\partial MsLoss}{\partial Z_p} = \frac{\partial MsLoss}{\partial \Lambda} \frac{\partial \Lambda}{\partial H} \frac{\partial H}{\partial g_\sigma} \frac{\partial g_\sigma}{\partial P_m} \frac{\partial P_m}{\partial Z_p} \quad (\text{E.1})$$

where $P_m = (Z(I))(m)$, m is a voxel at position (x_m, y_m, z_m) of the input image I , $\Lambda = (\lambda_1, \lambda_2, \lambda_3)$ and $MsLoss = \frac{1}{3M} \sum_{m=1}^M [(\lambda_{1_{P_m}} - \lambda_{1_{R_m}})^2 + (\lambda_{2_{P_m}} - \lambda_{2_{R_m}})^2 + (\lambda_{3_{P_m}} - \lambda_{3_{R_m}})^2]$.

For the sake of simplicity we define $V_m = g_\sigma * P_m$, and using the Magnus's theorem with an eigenvector $W_o = (w_{o1}, w_{o2}, w_{o3})'$ for each $\lambda_{o \in [1,3]}$, we have for a single voxel m (with M

equal to the number of voxels) with respect to a parameter Z_p :

$$\begin{aligned}
& \frac{\partial MsLoss}{\partial Z_p} = \\
& = \frac{\partial MsLoss}{\partial \Lambda} \frac{\partial \Lambda}{\partial H} \frac{\partial H}{\partial V_m} \frac{\partial V_m}{\partial Z_p} = \\
& = \frac{\partial MsLoss}{\partial \Lambda} \frac{\partial \Lambda}{\partial (vecH)'} \frac{\partial (vecH)'}{\partial V_m} \frac{\partial V_m}{\partial Z_p} = \\
& = \begin{bmatrix} \frac{\partial MsLoss}{\partial \lambda_1} & \frac{\partial MsLoss}{\partial \lambda_2} & \frac{\partial MsLoss}{\partial \lambda_3} \end{bmatrix} \times \begin{bmatrix} \frac{\partial \lambda_1}{\partial (vecH)'} \\ \frac{\partial \lambda_2}{\partial (vecH)'} \\ \frac{\partial \lambda_3}{\partial (vecH)'} \end{bmatrix} \times \begin{bmatrix} \frac{\partial h_{x,2}}{\partial (V_m)_1} & \dots & \frac{\partial h_{x,2}}{\partial (V_m)_M} \\ \dots & \dots & \dots \\ \frac{\partial h_{z,2}}{\partial (V_m)_1} & \dots & \frac{\partial h_{z,2}}{\partial (V_m)_M} \end{bmatrix} \times \begin{bmatrix} \frac{\partial (V_m)_1}{\partial Z_p} \\ \dots \\ \frac{\partial (V_m)_M}{\partial Z_p} \end{bmatrix} = \\
& = \begin{bmatrix} \frac{\partial MsLoss}{\partial \lambda_1} & \frac{\partial MsLoss}{\partial \lambda_2} & \frac{\partial MsLoss}{\partial \lambda_3} \end{bmatrix} \times \begin{bmatrix} W'_1 \otimes W'_1 \\ W'_2 \otimes W'_2 \\ W'_3 \otimes W'_3 \end{bmatrix} \times \begin{bmatrix} \frac{\partial h_{x,2}}{\partial (V_m)_1} & \dots & \frac{\partial h_{x,2}}{\partial (V_m)_M} \\ \dots & \dots & \dots \\ \frac{\partial h_{z,2}}{\partial (V_m)_1} & \dots & \frac{\partial h_{z,2}}{\partial (V_m)_M} \end{bmatrix} \times \begin{bmatrix} \frac{\partial (V_m)_1}{\partial Z_p} \\ \dots \\ \frac{\partial (V_m)_M}{\partial Z_p} \end{bmatrix}
\end{aligned}$$

$$\begin{aligned}
& \text{where } W'_o \otimes W'_o = \begin{bmatrix} w_{o1} & w_{o2} & w_{o3} \end{bmatrix} \otimes \begin{bmatrix} w_{o1} & w_{o2} & w_{o3} \end{bmatrix} = \\
& = \begin{bmatrix} w_{o1}w_{o1} & w_{o1}w_{o2} & w_{o1}w_{o3} & \dots & w_{o3}w_{o1} & w_{o3}w_{o2} & w_{o3}w_{o3} \end{bmatrix}
\end{aligned}$$

This matrix multiplication returns a single value, i.e. the gradient of $MsLoss$ with respect to the parameter Z_p , which will be multiplied by lr in order to update the parameter (e.g. weight) Z_p .

Comparison of different optimizers

In our method, we assessed also the performance of Adam [75] and Adagrad [34] optimizers with initial lr of 10^{-3} , taking inspiration from [78]. Results are shown in Table E.1, which show that the use of SGD with $lr=0.01$ leads to better performance.

Table E.1: Benchmarks using different optimizers and different lr on our proposed method using the baseline loss function: CE+Sof Dice with deep supervision. Patch size $32 \times 64 \times 64$. A: Arteries; V:Veins. Mean and standard deviation of the results are shown.

Technique		Structures					
Optimizer	lr	S	Dice Score [%] (\uparrow)	Precision [%] (\uparrow)	Recall [%] (\uparrow)	HD95 [mm] (\downarrow)	$\Delta\Lambda$ (\downarrow)
SGD	0.01	A	74.25 (4.24)	85.80 (11.25)	67.02 (7.45)	13.76 (17.15)	19.65 (2.04)
		V	58.67 (27.06)	82.73 (9.89)	51.10 (27.69)	16.17 (12.86)	19.77 (3.36)
SGD	0.001	A	73.39 (3.25)	86.54 (6.65)	64.48 (7.11)	9.23 (3.56)	20.35 (2.01)
		V	56.89 (28.44)	83.03 (8.45)	50.47 (28.78)	14.59 (11.67)	19.99 (3.40)
ADAM	0.01	A	69.97 (5.17)	79.07 (13.14)	64.89 (8.08)	20.21 (17.96)	19.91 (2.39)
		V	51.84 (24.70)	61.55 (17.61)	52.53 (28.82)	24.71 (24.36)	20.16 (3.36)
ADAM	0.001	A	74.32 (2.48)	85.88 (8.34)	66.52 (6.99)	9.48 (4.58)	19.92 (1.87)
		V	58.12 (26.24)	80.69 (12.85)	50.21 (27.59)	16.23 (17.87)	20.05 (3.63)
ADAGRAD	0.01	A	73.54 (4.83)	85.86 (11.79)	66.02 (7.64)	14.46 (17.13)	19.90 (2.38)
		V	57.40 (27.07)	83.69 (8.17)	50.31 (27.81)	16.05 (10.15)	19.76 (3.59)
ADAGRAD	0.001	A	60.34 (6.90)	53.75 (10.66)	70.77 (5.97)	16.11 (1.87)	18.67 (1.45)
		V	20.95 (14.76)	51.22 (27.09)	13.55 (10.11)	13.87 (3.56)	22.55 (2.98)

Parameters research

We conducted a parameters search for α , β and γ of Frangi's vesselness [38] and the best σ_{max} for ensuring zero gradient only in the main direction. The goal was to find the parameters for which the vesselness score F was greater than zero for all voxels of the target structures. Results are shown in the tables in Figure E.1.

BLOOD VESSELS %(F>0)	σ_{\max}				
α, β, γ	2	5	10	20	25
[0.1,0.1,2]	27%	60%	87%	97%	100%
[0.1,0.1,5]	27%	60%	87%	97%	100%
[0.1,0.1,15]	27%	60%	87%	97%	100%
[0.1,0.5,2]	27%	60%	87%	97%	100%
[0.1,0.5,5]	27%	60%	87%	97%	100%
[0.1,0.5,15]	27%	60%	87%	97%	100%
[0.5,0.1,2]	26%	59%	86%	96%	99%
[0.5,0.1,5]	26%	59%	86%	96%	99%
[0.5,0.1,15]	26%	59%	86%	96%	99%
[0.5,0.5,2]	26%	59%	86%	96%	99%
[0.5,0.5,5]	26%	59%	86%	96%	99%
[0.5,0.5,15]	26%	59%	86%	96%	99%

URETERS %(F>0)	σ_{\max}		
α, β, γ	5	15	25
[0.05,0.05,2]	41%	92%	100%
[0.1,0.1,2]	40%	91%	100%
[0.5,0.5,2]	39%	90%	99%

Figure E.1: Search of best parameters for σ_{max} of Gaussian filter and α , β and γ for Frangi's vesselness filter: we counted the percentage of voxels of target structures that have a Frangi vesselness greater than 0. The use of $\sigma_{max}=25$ and $\alpha=0.1$, $\beta=0.1$ and $\gamma=2$ allows values greater than zero for all voxels and thus distinguishing them from having no segmentation or from blob or plate structures. From the tables we can deduce that the most important value to set is σ_{max} to have a strong enough gradient given the larger cross sectional size of some vessels (e.g. aorta and cava vein) and of calyces (attachment of ureters to the kidneys). For the same reason, a smaller value of α also allows us to consider structures with a blob-like shape. As there are no plates or structures with little contrast, the weights of β and γ are irrelevant. Color code: from light to dark green as the percentage is better.

Implementation study of our proposed loss functions

Two other empirically studies were done using patches of size $32 \times 64 \times 64$: the first one to find the best implementation for Gaussian kernel application, while the second one to find the best w_{ms} of $MsLoss$. The choice of using little patches was made in order to make training faster and in order to verify the effectiveness of the proposed loss functions even when the ratio of foreground to background voxels is higher. However, even with the use of a method for ROI cropping, the use of such small patches greatly slows down the inference phase and it is not recommended for the implementation of pipelines for the creation of anatomical 3D models. Here, we made use of reference segmentations to extract the ROI where target structures are present. Results are respectively shown in Table E.2 and Table E.3, where the best methods are those later used as presented in Chapter 5. The application of multiple Gaussian kernels in all levels of resolution allows a better extraction of morphological information and thus better comparison between the predicted and reference structures. The first w_{ms} to weight $MsLoss$ was calculated in order to have values around 1, as Dice loss and Fvloss values are between 0 and 1. Due to the fact that $\Delta\Lambda$ for the baseline is approximately 20, we chose w_{ms} equal to 0.05. This value was then confirmed to be the best suited.

Table E.2: Results using 3D patches $32 \times 64 \times 64$ for different Gaussian kernel implementations. The values of $\sigma_{i \in [1,5]}$ are in the range $[1, \sigma_{max}]$ with a step of $\frac{\sigma_{max}}{5}$. We used as method for this study the nnU-Net ([61]) with U-Net as backbone and CE+Soft Dice with deep supervision as baseline loss function. Here w_{ms} is fixed at 0.05, σ_{max} at 25. A: Arteries; V: Veins. Mean and standard deviation of the results are shown.

Vesselness				Structures					
Loss	Deep Sup.	Post	σ_i	S	Dice Score [%] (\uparrow)	Precision [%] (\uparrow)	Recall [%] (\uparrow)	HD95 [mm] (\downarrow)	$\Delta\Lambda$ (\downarrow)
No	-	No	-	A	73.49 (3.87)	85.15 (10.80)	66.10 (7.21)	15.78 (18.19)	20.05 (2.19)
				V	57.07 (27.75)	84.73 (8.31)	48.79 (27.54)	17.19 (13.93)	20.01 (3.18)
MsLoss	No	No	[1,5]	A	74.17 (5.49)	83.73 (10.86)	68.11 (8.05)	14.21 (17.59)	19.12 (2.19)
				V	58.55 (29.16)	82.58 (6.47)	53.12 (30.26)	16.02 (12.77)	19.24 (3.57)
MsLoss	Yes	No	5	A	76.16 (5.16)	86.04 (10.88)	69.81 (7.79)	17.09 (16.86)	18.16 (1.89)
				V	57.78 (27.29)	82.81 (7.85)	51.43 (28.42)	21.15 (17.92)	19.63 (3.54)
MsLoss	Yes	No	[q+1] as Deep	A	74.14 (5.35)	80.78 (11.68)	70.23 (7.08)	14.13 (17.50)	18.71 (1.86)
				V	58.93 (26.89)	81.65 (8.12)	52.86 (28.39)	15.01 (11.23)	19.50 (3.44)
MsLoss	Yes	No	[1,(5-q)] as Deep	A	75.30 (4.92)	84.42 (11.34)	69.55 (7.51)	13.49 (17.97)	18.83 (2.09)
				V	60.68 (26.45)	80.94 (7.89)	55.16 (28.68)	16.18 (15.31)	18.99 (3.56)
MsLoss + FvLoss	No	No	[1,5]	A	76.84 (5.69)	80.04 (11.44)	75.60 (7.25)	15.69 (17.73)	17.23 (1.89)
				V	58.71 (27.05)	76.89 (9.16)	55.97 (30.28)	14.77 (10.91)	19.11 (3.26)
MsLoss + FvLoss	Yes	No	5	A	74.23 (5.60)	84.05 (12.91)	68.31 (7.10)	15.85 (17.53)	19.23 (1.71)
				V	58.41 (26.59)	82.44 (9.79)	50.39 (26.92)	16.63 (18.12)	20.17 (3.06)
MsLoss + FvLoss	Yes	No	[q+1] as Deep	A	75.07 (6.12)	83.96 (13.06)	69.89 (7.67)	15.66 (17.59)	18.73 (1.84)
				V	58.70 (25.83)	81.47 (11.19)	50.54 (26.26)	13.62 (11.61)	20.22 (3.10)
MsLoss + FvLoss	Yes	No	[1,(5-q)] as Deep	A	77.04 (7.18)	84.49 (13.47)	72.82 (7.65)	15.77 (17.58)	17.72 (1.97)
				V	58.04 (24.35)	81.30 (10.95)	50.54 (24.47)	15.44 (8.98)	17.83 (4.65)

Quantitative results for different implementations of tubular structures loss functions

Using the same small patches as before we tested the vesselness implementation used in [16] and [136] where eigenvalues are ordered by magnitude as in the original vesselness functions. Results are shown in Table E.4. As discussed in Section 5.2.2, direct comparison of vesselness score leads to worsening of results because two voxels belonging to vessels with different preferential directions (the predicted and reference ones) can have very similar vesselness

Table E.3: Results using 3D patches $32 \times 64 \times 64$ with different w_{ms} to weight $MsLoss$. For these tests we used our proposed method that differs on the oversampling technique used compared to nnU-Net [61]. We used U-Net as backbone and CE+Soft Dice+TsLoss with deep supervision as loss function. The proposed vesselness loss functions are applied also with deep supervision with the Gaussian kernel applied with $\sigma_{i \in [1, 5-q]}$ from 1 to σ_{max} with a step of $\frac{\sigma_{max}}{5}$, where q is the output resolution level (0 is the output at the same size of the input image). A: Arteries; V: Veins. Mean and standard deviation of the results are shown.

w_{ms}	S	Dice Score [%] (\uparrow)	Precision [%] (\uparrow)	Recall [%] (\uparrow)	HD95 [mm] (\downarrow)	$\Delta\Lambda$ (\downarrow)
0.01	A	77.19 (7.79)	79.56 (13.16)	76.86 (6.62)	15.08 (17.68)	17.07 (1.98)
	V	56.22 (26.92)	81.25 (8.84)	50.47 (29.15)	16.65 (12.79)	19.76 (3.63)
0.05	A	77.31 (4.42)	78.10 (12.48)	77.19 (6.64)	13.83 (17.39)	17.15 (1.81)
	V	60.20 (25.15)	76.38 (9.08)	57.22 (29.36)	14.58 (11.33)	19.11 (3.57)
0.1	A	72.79 (7.26)	70.61 (13.06)	77.72 (6.59)	15.69 (17.77)	19.92 (2.27)
	V	58.57 (25.08)	78.92 (8.07)	53.11 (27.85)	16.78 (15.32)	19.99 (3.50)
0.5	A	68.85 (6.83)	62.91 (12.19)	79.03 (6.77)	21.57 (19.77)	17.17 (1.99)
	V	58.87 (23.73)	72.76 (9.19)	55.58 (28.31)	19.36 (26.81)	18.98 (3.53)
1	A	51.72 (8.85)	40.64 (11.42)	75.87 (5.99)	31.72 (18.23)	16.16 (1.42)
	V	44.42 (15.31)	34.67 (11.52)	67.57 (26.91)	41.74 (25.13)	17.06 (3.62)

scores. Furthermore, the Recall analysis using patches $32 \times 64 \times 64$ is also shown in Figure E.2, which confirms what was stated in Section 5.4.

Table E.4: Results using 3D patches $32 \times 64 \times 64$ on combination of different loss functions. For these tests we used our proposed method that differs on the oversampling technique used compared to nnU-Net [61]. We used U-Net as backbone and CE+Soft Dice with deep supervision as baseline loss function. Here w_{ms} is fixed at 0.05, the proposed loss functions are applied also with deep supervision with the Gaussian kernel applied with $\sigma_{i \in [1, 5-q]}$ from 1 to σ_{max} with a step of $\frac{\sigma_{max}}{5}$, where q is the output resolution level (0 is the output at the same size of the input image). A: Arteries; V: Veins. Mean and standard deviation of the results are given. *The eigenvalues are ordered by magnitude as in the original vesselness functions.

Loss function used	S	Dice Score [%] (\uparrow)	Precision [%] (\uparrow)	Recall [%] (\uparrow)	HD95 [mm] (\downarrow)	$\Delta\Lambda$ (\downarrow)
No	A	74.25 (4.24)	85.80 (11.25)	67.02 (7.45)	13.76 (17.15)	19.65 (2.04)
	V	58.67 (27.06)	82.73 (9.89)	51.10 (27.69)	16.17 (12.86)	19.77 (3.36)
FvLoss* as Frangi SSVMD [16]	A	73.65 (5.16)	86.71 (11.72)	65.53 (7.10)	15.10 (17.03)	19.56 (1.96)
	V	58.81 (27.93)	80.41 (9.63)	53.40 (29.25)	14.42 (10.42)	19.42 (3.36)
FvLoss* as Jerman SSVMD [136]	A	73.13 (5.56)	85.37 (12.14)	65.63 (7.66)	14.71 (17.24)	20.03 (1.86)
	V	56.06 (27.40)	83.33 (8.05)	48.72 (27.74)	15.26 (11.83)	20.23 (3.32)
MsLoss (Λ in Ms ordered via W)	A	75.63 (4.57)	81.07 (11.64)	72.74 (7.55)	13.26 (17.54)	18.21 (1.88)
	V	59.73 (26.19)	79.81 (9.65)	53.73 (28.23)	14.71 (12.07)	19.23 (3.25)
MsLoss + FvLoss (Λ in Ms ordered via W)	A	77.31 (4.42)	78.10 (12.48)	77.19 (6.64)	13.83 (17.39)	17.15 (1.81)
	V	60.20 (25.15)	76.38 (9.08)	57.22 (29.36)	14.58 (11.33)	19.11 (3.57)

Other qualitative results

Other results are shown in Figure E.3 from the experiments done using smaller patches of size $32 \times 64 \times 64$.

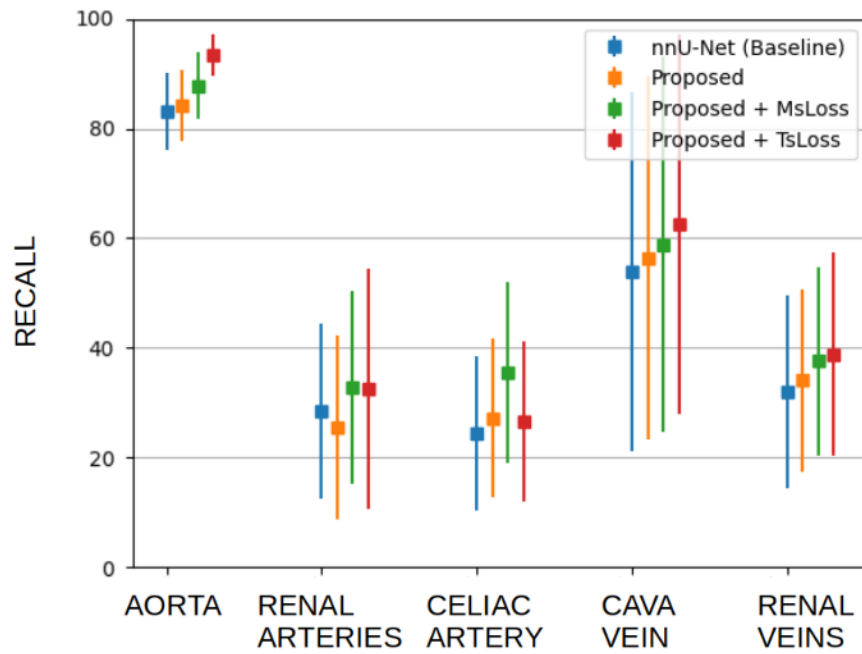


Figure E.2: Recall of the first row and last four rows of experiments in Table E.2 with patches $32 \times 64 \times 64$ for the different structures. Arteries are divided into aorta, renal arteries and celiac artery, while veins in cava vein and renal veins.

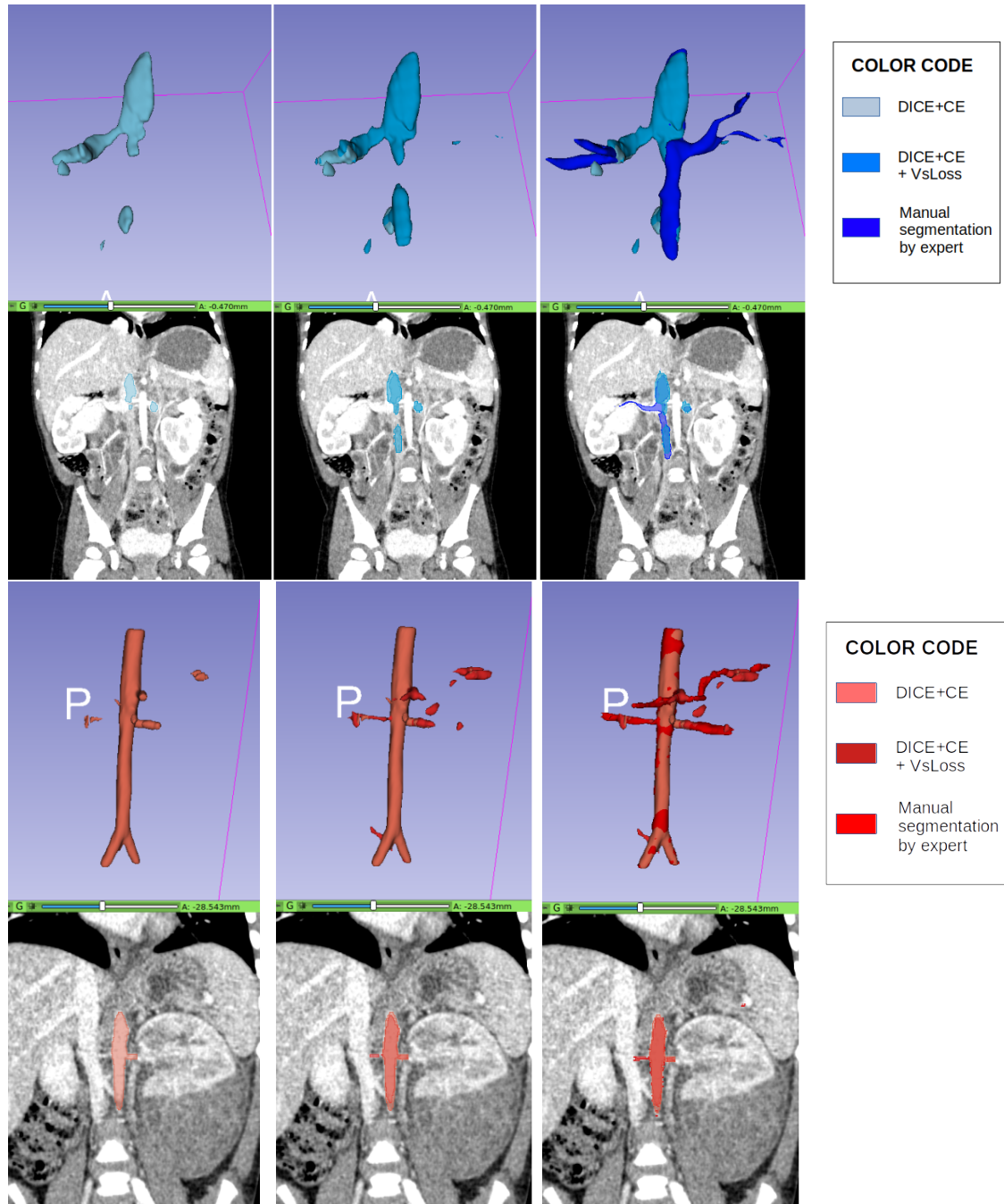


Figure E.3: Example of segmentation for veins (top) and arteries (bottom) on two difficult cases. The top ceCT image presents a strong heterogeneity in the cava vein due to the tumor presence. The bottom ceCT image presents renal arteries with a very few voxels.

Appendix F

Details of the “Renal Anatomy Segmentation for ceCT” module

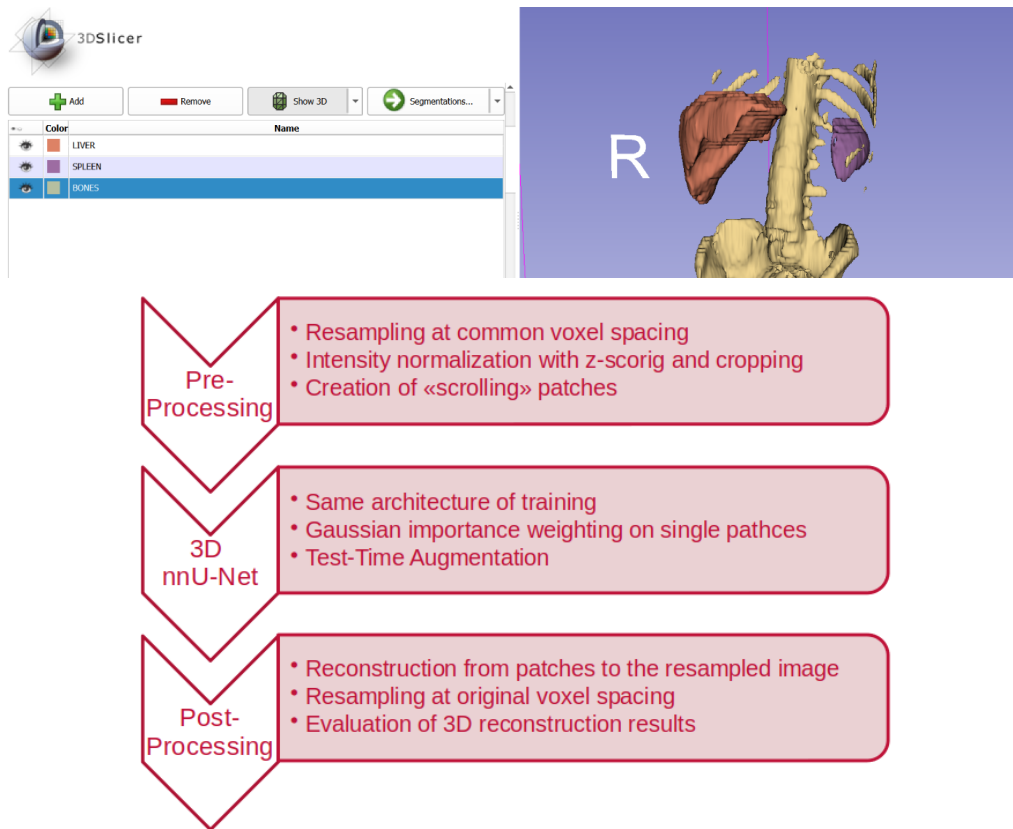


Figure F.1: Top: example of step 2 of Selection and Segmentation section: automatic segmentation of bones, liver and spleen. Bottom: inference phase used in our 3DSlicer [37] plug-in.

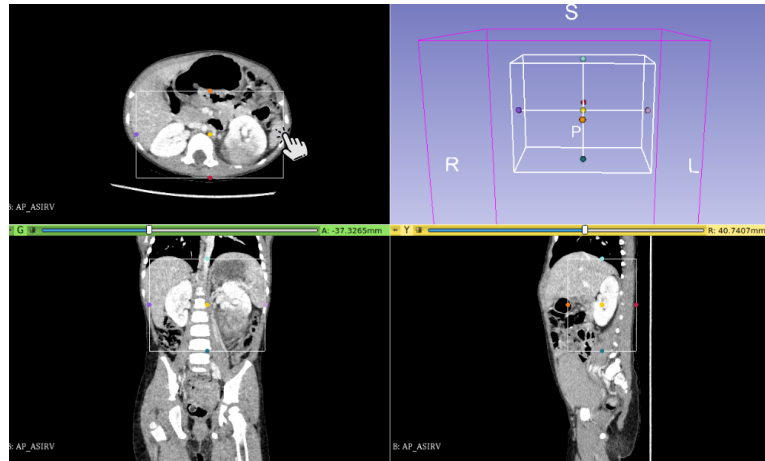


Figure F.2: Example of step 3 of Selection and Segmentation section: creation and interaction of the AnnotationROI.

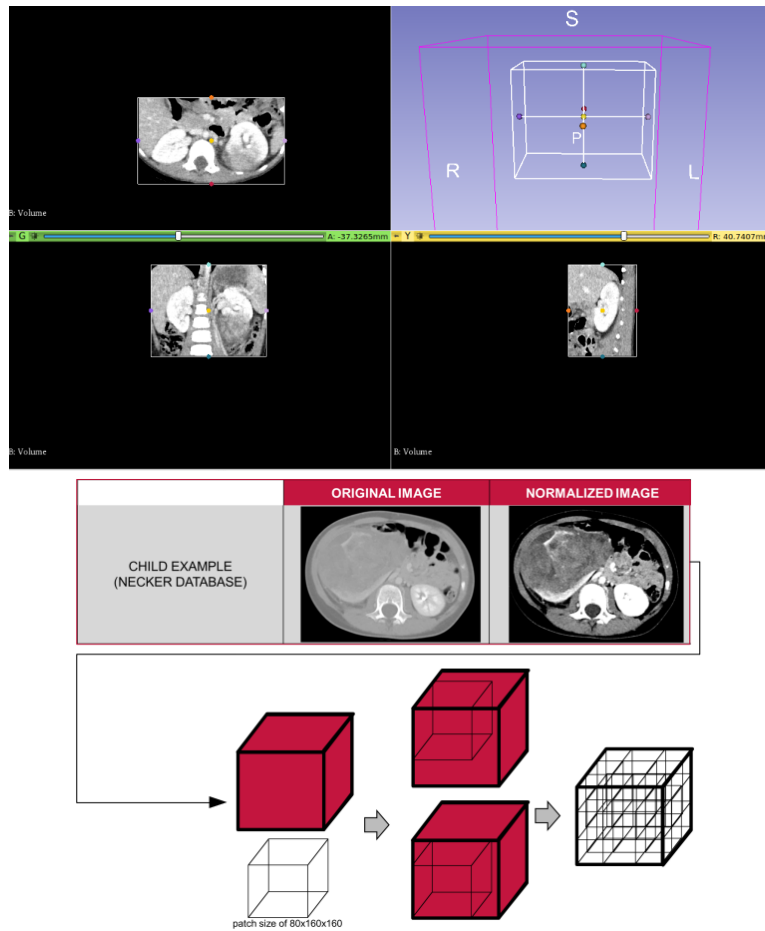


Figure F.3: Top: example of step 4 of Selection and Segmentation section: creation and interaction of the AnnotationROI. Bottom: visual explanation of the pre-processing.

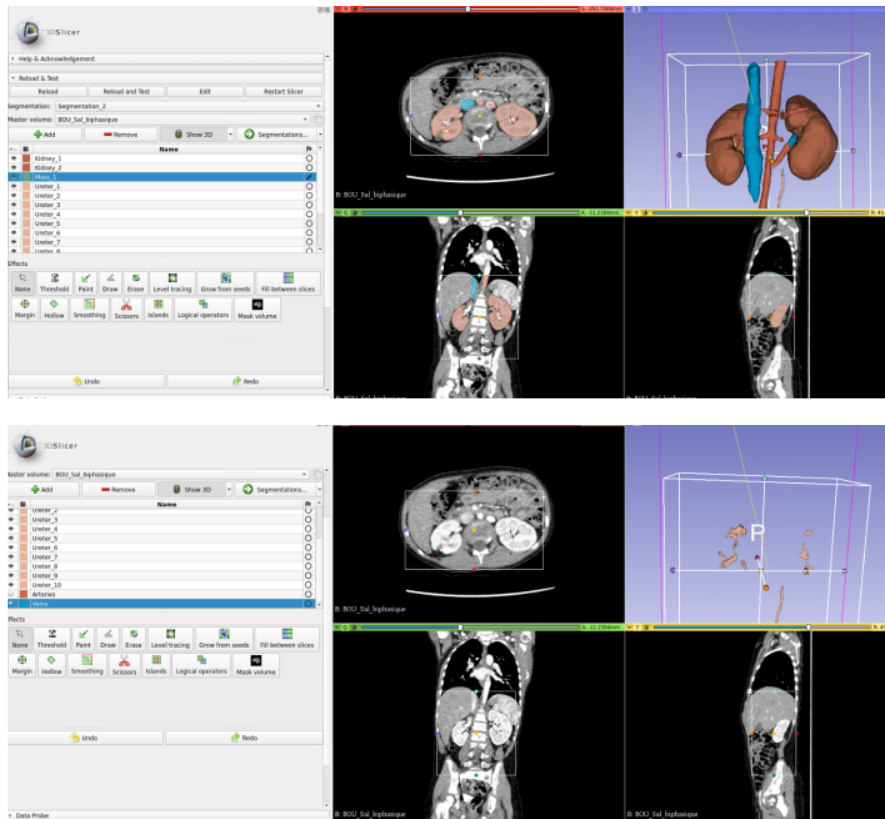


Figure F.4: Top: example of step 5 of Selection and Segmentation section: automatic segmentation of kidneys, tumors, arteries, veins and ureters. Bottom: focus on ureters.

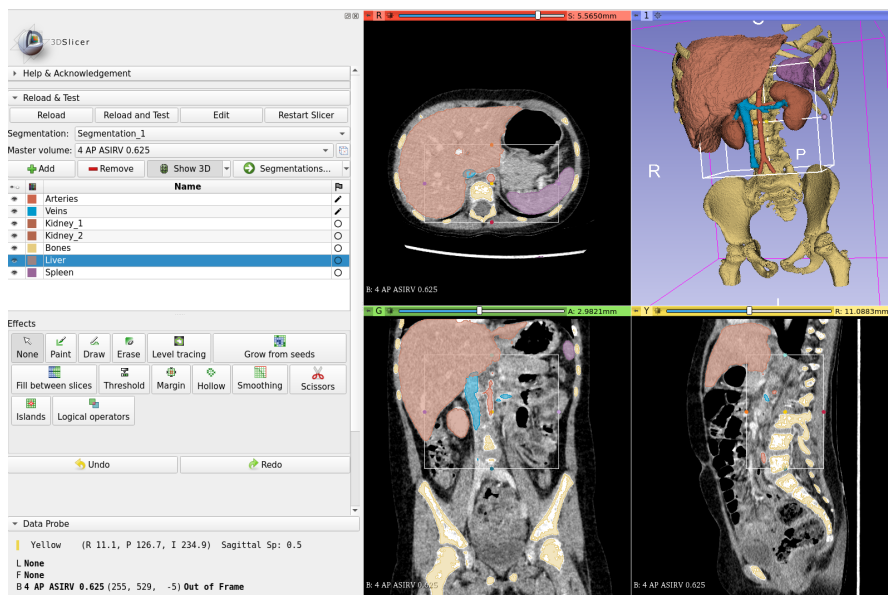


Figure F.5: Example of final automatic 3D segmentation obtained with our plug-in.

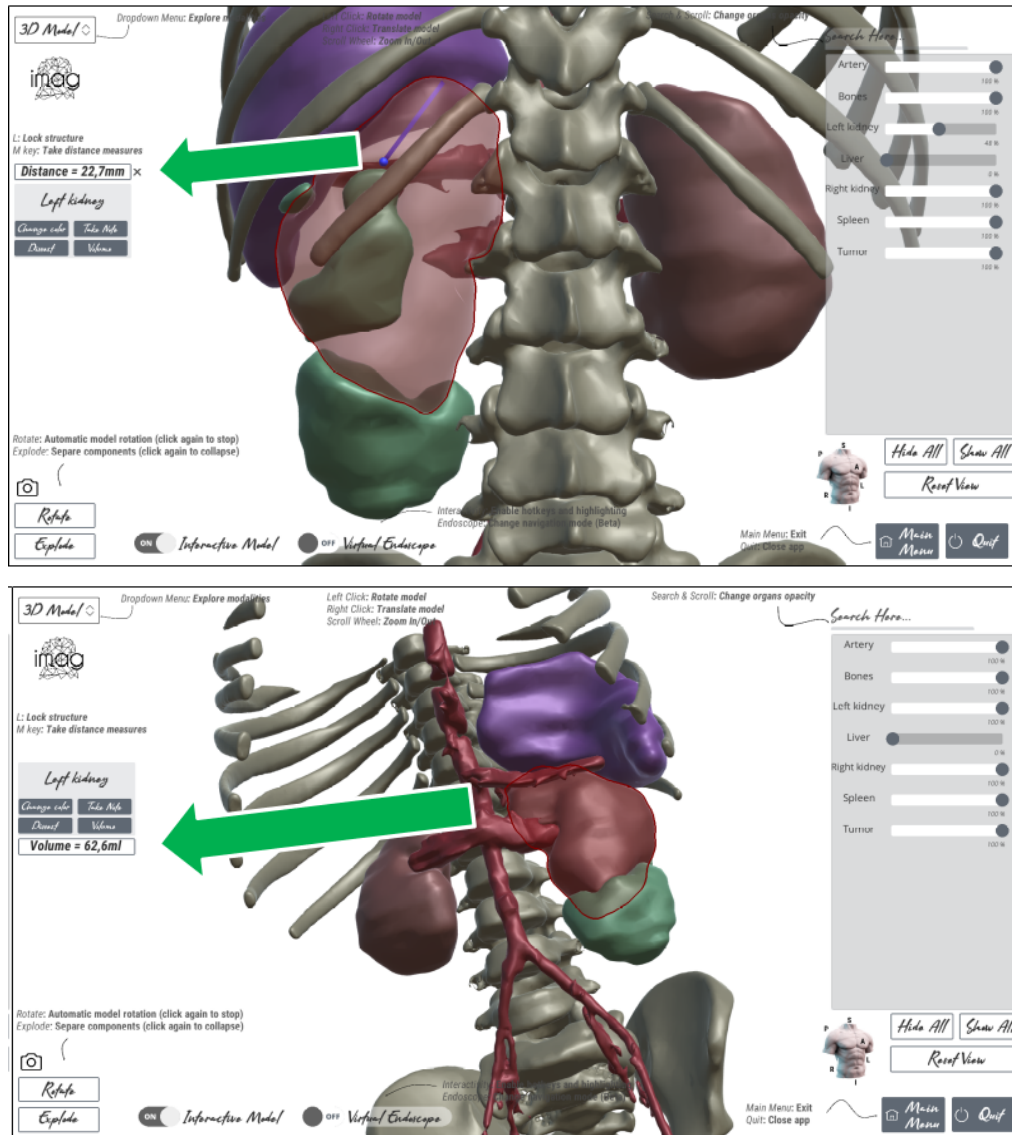


Figure F.6: Some images of the software tool specifically designed at the IMAG2 lab of Necker hospital for visualization and interaction. Top: distance calculation in *mm* between two user-defined points. Bottom: volume in *ml* of the selected structure.


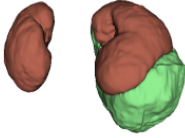

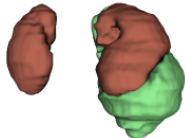
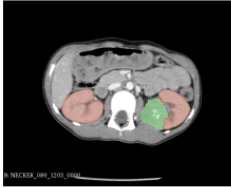
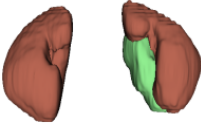

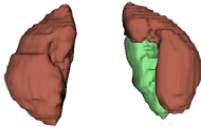
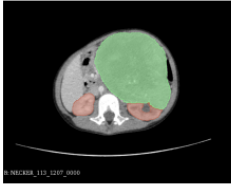
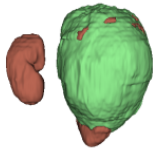
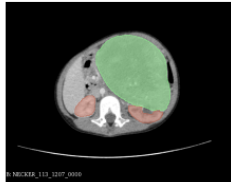
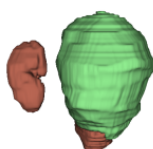
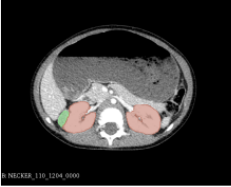
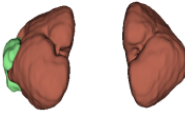

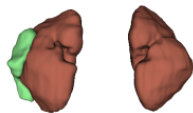
Example Dice Score	Automatic Segmentation 2D slice – 3D model	Manual Segmentation 2D slice – 3D model
BEST RESULT Kidney = 94.7% Tumor = 93.8%	 	 
AVERAGE RESULT Kidney = 92.8% Tumor = 88.2%	 	 
WORST KIDNEY RESULT Kidney = 82.9% Tumor = 95.3%	 	 
WORST TUMOR RESULT Kidney = 93.9% Tumor = 54.7%	 	 

Figure F.7: Best, average and worst results for kidney and tumor segmentation using the best method found, namely the one presented in Table 6.2, trained using our 3D nnU-Net [61] implementation using as input both real ceCT images and synthetic CT images generated with the method proposed in Chapter 4. These results are the ones used in our plug-in.

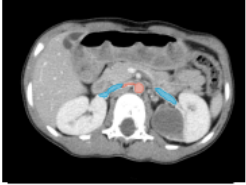
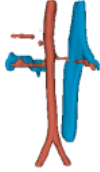
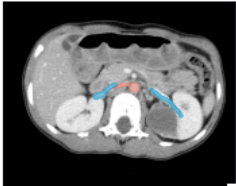
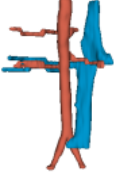


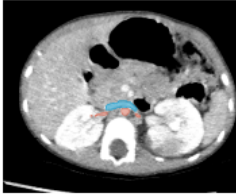
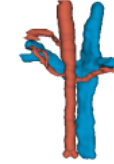
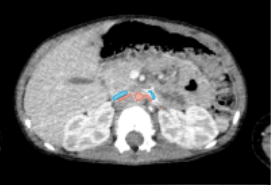

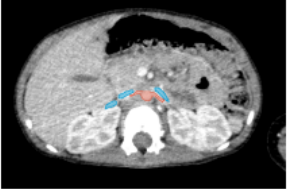
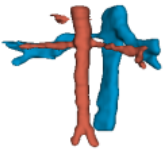
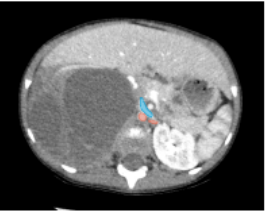

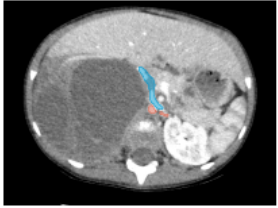

Example Dice Score($\Delta\Lambda$)	Automatic Segmentation 2D slice – 3D model		Manual Segmentation 2D slice – 3D model	
BEST RESULT Arteries = 85.6% ($\Delta\Lambda = 18.3$) Veins = 81.6% ($\Delta\Lambda = 15.1$)				
AVERAGE RESULT Arteries = 74.4% ($\Delta\Lambda = 20.8$) Veins = 55.9% ($\Delta\Lambda = 22.9$)				
WORST ARTERIES RESULT Arteries = 67.8% ($\Delta\Lambda = 20.9$) Veins = 67.8% ($\Delta\Lambda = 22.2$)				
WORST VEINS RESULT Arteries = 72.3% ($\Delta\Lambda = 20.4$) Veins = 22.9 ($\Delta\Lambda = 20.3$)				

Figure F.8: Best, average and worst results for arteries and veins segmentation using the best method found, namely the one presented in Table 6.2, trained using our 3D nnU-Net [61] implementation with the use of oversampling method and the Tubular structures Loss proposed in Chapter 5, using as input both real ceCT images and synthetic CT images generated with the method proposed in Chapter 4. These results are the ones used in our plug-in.




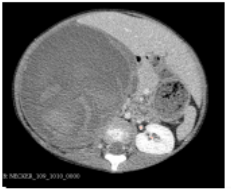
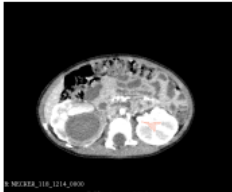
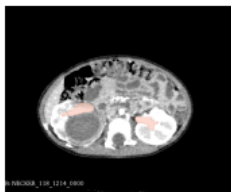
Example Dice Score($\Delta\Lambda$)	Automatic Segmentation 2D slice – 3D model	Manual Segmentation 2D slice – 3D model
BEST RESULT Ureters = 78.9% ($\Delta\Lambda = 22.2$)		
AVERAGE RESULT Ureters = 62.9% ($\Delta\Lambda = 24.18$)		
WORST RESULT Ureters = 19.7% ($\Delta\Lambda = 9.9$)		

Figure F.9: Best, average and worst results for ureters segmentation using the best method found, namely the one presented in Table 6.2, trained using our 3D nnU-Net [61] implementation with the use of oversampling method and the Morphological similarity Loss proposed in Chapter 5, using as input both real ceCT images and synthetic CT images generated with the method proposed in Chapter 4. These results are the ones used in our plug-in.

Bibliography

- [1] Alborz Amir-Khalili, Jean-Marc Peyrat, Julien Abinahed, Osama Al-Alao, Abdulla Al-Ansari, Ghassan Hamarneh, and Rafeef Abugharbieh. Auto localization and segmentation of occluded vessels in robot-assisted partial nephrectomy. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, volume 17, pages 407–414, 2014.
- [2] Martin Arjovsky, Soumith Chintala, and Lottou Bottou. Wasserstein GAN. In *International Conference on Machine Learning (ICML)*, volume 70, pages 214–223, 2017.
- [3] Karim Armanious, Chenming Jiang, Sherif Abdulatif, Thomas Küstner, Sergios Gatidis, and Bin Yang. Unsupervised medical image translation using Cycle-MedGAN. In *European Signal Processing Conference (EUSIPCO)*, pages 1–5, 2019.
- [4] Guha Balakrishnan, Amy Zhao, Mert Sabuncu, John Guttag, and Adrian Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.
- [5] Christian Bauer and Horst Bischof. A novel approach for detection of tubular objects and its application to medical image analysis. In *DAGM Symp. Pattern Recognition*, volume 5096, pages 163–172, 2008.
- [6] Michael Baumgartner, Paul F Jäger, Fabian Isensee, and Klaus H Maier-Hein. nnDetection: a self-configuring method for medical object detection. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 530–539, 2021.
- [7] Cam Bermudez, Justin Blaber, Samuel Remedios, Jess Reynolds, Catherine Lebel, Maureen Mchugo, Stephan Heckers, Yuankai Huo, and Bennett Landman. Generalizing deep whole brain segmentation for pediatric and post-contrast MRI with augmented transfer learning. In *SPIE Medical Imaging*, volume 11313, 2020.
- [8] Christoph Boeddeker, Patrick Hanebrink, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach. On the computation of complex-valued gradients with application to statistically optimum beamforming. *ArXiv*, abs/1701.00392, 2017.
- [9] Alexandre Bone, Maxime Louis, Olivier Colliot, and Stanley Durrleman. Learning low-dimensional representations of shape data sets with diffeomorphic autoencoders. In *Intelligent Platform Management Interface (IPMI)*, volume 11492, pages 195–207, 2019.
- [10] F L Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

- [11] G Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [12] Carmela Brilliantino, Eugenio Rossi, Rocco Minelli, and Elio Bignardi. Current role of imaging in the management of children with Wilms tumor according to the new UMBRELLA protocol. *Translational Medicine*, pages 30–39, 2019.
- [13] Katarzyna Bugajska, Andrzej Skalski, Janusz Gajda, and Tomasz Drewniak. The renal vessel segmentation for facilitation of partial nephrectomy. *Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 50–55, 2015.
- [14] Alexandre Bône, S. Ammari, Jean-Philippe Lamarque, Mickael Elhaik, Emilie Chouzenoux, François Nicolas, Philippe Robert, Corinne Balleyguier, Nathalie Lassau, and Marc-Michel Rohe. Contrast-Enhanced brain MRI synthesis with deep learning: Key input modalities and asymptotic performance. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1159–1163, 2021.
- [15] Francesco Caliva, Claudia Iriondo, Alejandro Morales Martinez, Sharmila Majumdar, and Valentina Pedoia. Distance map loss penalty term for semantic segmentation. In *Medical Imaging with Deep Learning (MIDL)*, 2019.
- [16] Kunlin Cao, Kaifang Du, Kai Ding, Joseph Reinhardt, and Gary Christensen. Regularized nonrigid registration of lung CT images by preserving tissue volume and vesselness measure. In *Medical Image Analysis for the Clinic: A Grand Challenge*, pages 43–54, 2010.
- [17] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016.
- [18] Liang-Chieh Cehn, 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, 2018.
- [19] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. PairedCycleGAN: Asymmetric style transfer for applying and removing makeup. In *IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 40–48, 2018.
- [20] Michael Chappell. *Principles of Medical Imaging for Engineers*. Springer, 2019.
- [21] Yann Chaussy, Lorédane Vieille, Elise Lacroix, Marion Lenoir, Florent Marie, Lisa Corbat, Julien Henriët, and Frédéric Auber. 3D reconstruction of Wilms' tumor and kidneys in children: Variability, usefulness and constraints. *Pediatric Urology*, 16(6):830.e1–830.e8, 2020.
- [22] Mingqing Chen, Kunlin Cao, Yefeng Zheng, and R Afredo C Siochi. Motion-compensated mega-voltage cone beam CT using the deformation derived directly from 2D projection images. *IEEE Transactions on Medical Imaging*, 32, 2012.

- [23] Xu Chen, Chunfeng Lian, Li Wang, Hannah Deng, Tianshu Kuang, Steve Fung, Jaime Gateno, and Pew-Thian Yap. Diverse data augmentation for learning image segmentation with cross-modality annotations. *Medical Image Analysis*, 71:102060, 2021.
- [24] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and F Prior. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *Digital Imaging*, 26(6):1045–1057, 2013.
- [25] Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 529–536, 2018.
- [26] Taco S Cohen and Max Welling. Group equivariant convolutional networks. In *International Conference on Machine Learning (ICML)*, volume 48, pages 2990–2999, 2016.
- [27] Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. ResViT: Residual vision transformers for multi-modal medical image synthesis. *IEEE Transactions on Medical Imaging*, PP:1–1, 2022.
- [28] Vien Ngoc Dang, Francesco Galati, Rosa Cortese, Giuseppe Di Giacomo, Viola Marconetto, Prateek Mathur, Karim Lekadir, Marco Lorenzi, Ferran Prados, and Maria A Zuluaga. Vessel-CAPTCHA: an efficient learning framework for vessel annotation and segmentation. *Medical Image Analysis*, 75:102263, 2022.
- [29] Salman Ul Hassan Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Çukur. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. *IEEE Transactions on Medical Imaging*, 38(10):2375–2388, 2019.
- [30] Laura E Diment, Mark S Thompson, and Jeroen H M Bergmann. Clinical efficacy and effectiveness of 3d printing: a systematic review. *BMJ Open*, 7(12):e016891, 2017.
- [31] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.
- [32] Qi Dou, Hao Chen, Yueming Jin, Lequan Yu, Jing Qin, and Pheng-Ann Heng. 3D deeply supervised network for automatic liver segmentation from CT volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 149–157, 2016.
- [33] Kaifang Du, Joseph M Reinhardt, Gary E Christensen, Kai Ding, and Bayouthm John E. Respiratory effort correction strategies to improve the reproducibility of lung expansion measurements. *Medical Physics*, 40(12):123504, 2013.
- [34] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Machine Learning Research*, 12:2121–2159, 2011.
- [35] Okechukwu Felix Erundu. Challenges and peculiarities of paediatric imaging. *Medical Imaging in Clinical Practice*, 23:23–35, 2013.

- [36] Gang Fan, Jun Li, Mingfeng Li, Mingji Ye, Xiaming Pei, Feiping Li, Shuai Zhu, Han Weiqin, Xiao Zhou, and Yu Xie. Three-dimensional physical model-assisted planning and navigation for laparoscopic partial nephrectomy in patients with endophytic renal tumors. *Scientific Reports*, 8(1):582–587, 2018.
- [37] Andrey Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, and Julien Finet. 3D Slicer as an image computing platform for the quantitative imaging network. *Magnetic Resonance Imaging*, (30(9)):1323–1341, 2012.
- [38] Ro Frangi, W.J. Niessen, Koen Vincken, and Max Viergever. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume 1496, 2000.
- [39] Hiroshi Fukui, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 10697–10706, 2019.
- [40] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Machine Learning Research*, 17(1):2096–2130, 2016.
- [41] Antonio Garcia-Uceda Juarez, Raghavendra Selvan, Zaigham Saghir, and Marleen de Bruijne. A joint 3D UNet-Graph Neural Network-based method for airway segmentation from chest CTs. In *Medical Image Computing and Computer Assisted Intervention (MICCAI) MLMI Challenge*, pages 583–591, 2019.
- [42] Yunhao Ge, Dongming Wei, Jon Xue, Qian Wang, Xiang Zhou, Yiqiang Zhan, and Shu Liao. Unpaired MR to CT synthesis with explicit structural constrained adversarial learning. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2019.
- [43] Yasmeen George. A coarse-to-fine 3D U-Net network for semantic segmentation of kidney CT scans. In *Medical Image Computing and Computer Assisted Intervention (MICCAI) KiTS Challenge*, pages 137–142, 2021.
- [44] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587, 2014.
- [45] Oscar Girón-Vallejo, Dario García-Calderón, and Ramon Ruiz-Pruneda. Three-dimensional printed model of bilateral Wilms tumor: A useful tool for planning nephron sparing surgery. *Pediatric Blood Cancer*, 65(4), 2018.
- [46] A Golts, D Khapun, D Shats, et al. An ensemble of 3D U-Net based models for segmentation of kidney and masses in CT scans. In *Medical Image Computing and Computer Assisted Intervention (MICCAI) KiTS Challenge*, pages 103–115, 2021.
- [47] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014.

- [48] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5767–5777, 2017.
- [49] Ali Hatamizadeh, Ducheng Yang, Holger R Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. In *Winter Conference on Applications of Computer Vision (WACV)*, pages 1748–1758, 2022.
- [50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 770–778, 2016.
- [51] Yuting He, Guanyu Yang, Jian Yang, Yang Chen, Youyong Kong, Jiasong Wu, Lijun Tang, Xiaomei Zhu, Jean-Louis Dillenseger, Pengfei Shao, Shaobo Zhang, Huazhong Shu, Jean-Louis Coatrieux, and Shuo Li. Dense biased networks with deep priori anatomy and hard region adaptation: Semi-supervised learning for fine renal artery segmentation. *Medical Image Analysis*, 63, 2020.
- [52] N Heller, N Papanikolopoulos, and C Weight. 2021 Kidney and Kidney Tumor Segmentation Challenge in 24th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI). <https://doi.org/10.5281/zenodo.3714972>, Zenodo version 0.19.6, 2020.
- [53] Nicholas Heller, Fabian Isensee, Klaus Maier-Hein, Xiaoshuai Hou, Chunmei Xie, Fengyi Li, Yang Nan, Guangrui Mu, Zhiyong Lin, Miofei Han, Guang Yao, Yaozong Gao, Yao Zhang, Yixin Wang, Feng Hou, Jiawei Yang, Guangwei Xiong, Jiang Tian, Cheng Zhong, and Christopher Weight. The state of the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. *Medical Image Analysis*, 67:101821, 2021.
- [54] Katarzyna Heryan, Dominik Choragwicki, Marek Sandheim, Jacek Jakubowski, and Tomasz Drewniak. Renal vessels segmentation for preoperative planning in percutaneous nephrolithotomy. In *Imaging Systems and Techniques (IST)*, pages 1–6, 2018.
- [55] Hesam M Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Digital Imaging*, 32(4):582–596, 2019.
- [56] Roland Hess. *Blender Foundations: The Essential Guide to Learning Blender 2.6*. Focal Press, 2010.
- [57] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6626–6637, 2017.
- [58] Marry Heuvel-Eibrink, Janna Hol, Kathy Pritchard-Jones, Harm Tinteren, Rhoikos Furtwängler, Arnould Verschuur, Gordan Vujanic, Ivo Leuschner, Jesper Brok, Christian Rübe, Anne Smets, Geert Janssens, Jan Godzinski, Gema Ramirez-Villar, Beatriz

- Camargo, Heidi Segers, Paola Collini, Manfred Gessler, Christophe Bergeron, and International SIOP-RTSG. Position paper: Rationale for the treatment of Wilms tumour in the UMBRELLA SIOP-RTSG 2016 protocol. *Nature Reviews Urology*, (14(12)):743–752, 2017.
- [59] E Hyde, L Berger, Navin Ramachandran, Archie Hughes-Hallett, N Pavithran, Maxine Tran, S Ourselin, Axel Bex, and F Mumtaz. Interactive virtual 3D models of renal cancer patient anatomies alter partial nephrectomy surgical planning decisions and increase surgeon confidence compared to volume-rendered images. *International Journal for Computer Assisted Radiology and Surgery (IJCARs)*, 14:723–732, 2019.
- [60] Sabine Irtan, Erik Hervieux, Hélène Boutroux, François Becmeur, Hubert Ducou-le-Pointe, Guy Leverger, and Georges Audry. Preoperative 3D reconstruction images for paediatric tumours: Advantages and drawbacks. *Pediatric Blood Cancer*, 68(1):e28670, 2021.
- [61] Fabian Isensee, Paul Jaeger, Simon Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:1–9, 2021.
- [62] Fabian Isensee, Paul Jäger, Jakob Wasserthal, David Zimmerer, Jens Petersen, Simon Kohl, Justus Schock, Andre Klein, Tobias Roß, Sebastian Wirkert, Peter Neher, Stefan Dinkelacker, Gregor Köhler, and Klaus Maier-Hein. Batchgenerators - a python framework for data augmentation. <https://doi.org/10.5281/zenodo.3632567>, Zenodo version 0.19.6, 2020.
- [63] Fabian Isensee and Klaus Maier-Hein. An attempt at beating the 3D U-Net. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) KiTS challenge*, 2019.
- [64] Amirul Islam, Sen Jia, and Neil D B Bruce. How much position information do convolutional neural networks encode? In *International Conference on Learning Representations (ICLR)*, 2020.
- [65] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial Transformer Networks. In *Neural Information Processing Systems (NeurIPS)*, 2015.
- [66] Tim Jerman, Franjo Pernus, Bostjan Likar, and Ziga Spiclin. Enhancement of vascular structures in 3D and 2D angiographic images. *IEEE Transaction of Medical Imaging*, 35:2107–2118, 2016.
- [67] Jinmeng Jia, Zhongxin An, Yue Ming, Yongli Guo, Wei Li, Xin Li, Yunxiang Liang, Dongming Guo, Jun Tai, Geng Chen, Yaqiong Jin, Zhimei Liu, Xin Ni, and Tielu Shi. PedAM: a database for pediatric disease annotation and medicine. *Nucleic Acids Research*, 46(D1):D977–D983, 2017.
- [68] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. TransGAN: Two pure transformers can make one strong GAN, and that can scale up. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

- [69] Lukasz Kaiser, Aidan Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017.
- [70] Eunhee Kang, Hyun Jung Koo, Dong Hyun Yang, Joon Bum Seo, and Jong Chul Ye. Cycle-consistent adversarial denoising network for multiphase coronary CT angiography. *Medical physics*, 46(2):550–562, 2019.
- [71] S Karthika and M Durgadevi. Generative adversarial network (GAN): a general review on different variants of GAN and applications. In *International Conference on Computational and Experimental Engineering and Sciences (ICCES)*, pages 1–8, 2021.
- [72] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-GAT-IT: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations (ICLR)*, pages 1–19, 2020.
- [73] Soohyun Kim, Jongbeom Baek, Jihye Park, Gyeongnyeon Kim, and Seungryong Kim. InstaFormer: Instance-aware image-to-image translation with transformer. In *Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [74] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [75] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
- [76] Jens Kleesiek, Jan Morshuis, Fabian Isensee, Katerina Deike-Hofmann, Daniel Paech, Philipp Kickingereder, Ullrich Köthe, Carsten Rother, Michael Forsting, Wolfgang Wick, Martin Bendszus, Heinz-Peter Schlemmer, and Alexander Radbruch. Can virtual contrast enhancement in brain MRI replace gadolinium?: A feasibility study. *Investigative Radiology*, 54:1, 2019.
- [77] Giammarco La Barbera, Isabelle Bloch, Gonzalo Barraza, Catherine Adasbaum, and Pietro Gori. Robust segmentation of corpus callosum in multi-scanner pediatric T1-w MRI using transfer learning. In *Organization for Human Brain Mapping (OHBM)*, 2019.
- [78] Mounir Lahlouh, Yasmina Chenoune, Raphael Blanc, Jérôme Szewczyk, and Nicolas Passat. Aortic arch anatomy characterization from MRA: A CNN-based segmentation approach. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022.
- [79] Jonas Lamy, Odyssée Merveille, Bertrand Kerautret, and Nicolas Passat. A benchmark framework for multi-region analysis of vesselness filters. *IEEE Transactions on Medical Imaging*, 41(12):3649–3662, 2022.
- [80] Jonas Lamy, Odyssée Merveille, Bertrand Kerautret, Nicolas Passat, and Antoine Vacavant. Vesselness filters: A survey with benchmarks applied to liver imaging. In *International Conference on Pattern Recognition (ICPR)*, pages 3528–3535, 2020.

- [81] Jonas Lamy, Thibault Pelletier, Guillaume Lienemann, Benoît Magnin, Bertrand Kerautret, Nicolas Passat, Julien Finet, and Antoine Vacavant. The 3D Slicer RVXLiverSegmentation plug-in for interactive liver anatomy reconstruction from medical images. *Open Source Software*, 7(73):3920.
- [82] B Landman, Z Xu, J Eugenio Igelsias, M Styner, T Langerak, and A Klein. Data from CT-Abdomen - MICCAI multi-atlas labeling beyond the cranial vault-Workshop and Challenge, 2015.
- [83] Max W K Law and Albert C S Chung. Three dimensional curvilinear structure detection using optimally oriented flux. In *European Conference on Computer Vision (ECCV)*, pages 368–382, 2008.
- [84] David Lesage, Elsa Angelini, Isabelle Bloch, and Gareth Funka-Lea. A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes. *Medical Image Analysis*, 13:819–845, 2009.
- [85] Junning Li, Pechin Lo, Ahmed Taha, Hang Wu, and Tao Zhao. Segmentation of renal structures for image-guided surgery. In *Medical Image Computing and Computer Assisted Interventions (MICCAI)*, pages 454–462, 2018.
- [86] Hanxiao Liu, Zihang Dai, David R So, and Quoc V Le. Pay attention to MLPs. In *Neural Information Processing Systems (NeurIPS)*, 2021.
- [87] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 700–708, 2017.
- [88] Zhaung Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Darrell Trevor, and Saining Xie. A ConvNet for the 2020s. In *Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022.
- [89] C Lorenz, I Carlsen, Thorsten Buzug, C Fassnacht, and J Weese. Multi-scale line segmentation with automatic estimation of width, contrast and tangential direction in 2D and 3D medical images. In *CVRMed-MRCAS*, pages 233–242, 1997.
- [90] Jun Ma, Zhan Wei, Yiwen Zhang, Yixin Wang, Rongfei Lv, Cheng Zhu, Gaoxiang Chen, Jianan Liu, Chao Peng, Lei Wang, Yunpeng Wang, and Jianan Chen. How distance transform maps boost segmentation CNNs: An empirical study. In *Medical Imaging with Deep Learning (MIDL)*, volume 121, pages 479–492, 2020.
- [91] Jan R Magnus. *Econometric Theory*, volume 1, chapter On Differentiating Eigenvalues and Eigenvectors, pages 179–191. Cambridge University Press, 1985.
- [92] Marcin Majos, Agata Majos, Michal Polgaj, Konrad Scyzmczyk, Jakub Chrostowski, and Ludomir Stefanczyk. Diameters of arteries supplying horseshoe kidneys and the level they branch off their parental vessels: A CT-angiographic study. *Clinical Medicine*, 8, 2019.

- [93] Savina Mannarino, Patrizia Bulzomì, Alessia Claudia Codazzi, Gaetana Anna Rispoli, Carmine Tinelli, Annalisa De Silvestri, Federica Manzoni, and Silvia Chiapedi. Inferior vena cava, abdominal aorta, and IVC-to-aorta ratio in healthy Caucasian children: Ultrasound Z-scores according to BSA and age. *Cardiology*, 74(4):388–393, 2019.
- [94] Jacques Marescaux and Michele Diana. Inventing the future of surgery. *Surgery*, 39(3):615–622, 2015.
- [95] Florent Marie, Lisa Corbat, Yann Chaussy, Thibault Delavelle, Julien Henriët, 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.
- [96] Kasper Marstal, Floris Berendsen, Marius Staring, and Stefan Klein. SimpleElastix: A user-friendly, multi-lingual library for medical image registration. In *International Workshop on Biomedical Image Registration (WBIR)*, 2016.
- [97] David Mattes, David R Haynor, Hubert Vesselle, Thomas K Lewellyn, and William Eubank. Nonrigid multimodality image registration. In *SPIE Medical Imaging*, pages 1609–1620, 2001.
- [98] Dirk Merkel. Docker: lightweight linux containers for consistent development and deployment. *Linux journal*, 2014(239):2, 2014.
- [99] Odyssee Merveille, Hugues Talbot, Laurent Najman, and Nicolas Passat. Curvilinear structure analysis by ranking the orientation responses of path operators. *IEEE Transaction on Pattern Analysis*, 40:304–317, 2018.
- [100] C Michiels, E Jambon, and J C Bernhard. Measurement of the accuracy of 3D-Printed medical models to be used for robot-assisted partial nephrectomy. *Am J Roentgenol*, 213(3):626–631, 2019.
- [101] Elena Giulia Milano, Caudio Capelli, Jo Wray, Benedetta Biffi, Sofie Layton, Matthew Lee, Massimo Caputo, Andrew M Taylor, Silvia Schievano, and Giovanni Biglino. Current and future applications of 3D printing in congenital cardiology and cardiac surgery. *British Journal of Radiology*, 92(1094):20180389, 2019.
- [102] Sara Moccia, Elena De Momi, Leonardo Mattos, and Sara El Hadji. Blood vessel segmentation algorithms — Review of methods, datasets and evaluation metrics. *Computer Methods and Programs in Biomedicine*, 158:71–91, 2018.
- [103] Nikita Moriakov, Jonas Adler, and Jonas Teuwen. Kernel of CycleGAN as a principle homogeneous space. In *International Conference on Learning Representations (ICLR)*, 2020.
- [104] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *International Conference on Learning Representations (ICLR)*, 2017.

- [105] Catherine Owens, Herve Brisse, Øystein Olsen, Joanna Begent, and Anne Smets. Bilateral disease and new trends in Wilms tumour. *Pediatric Radiology*, (38(1)):30–39, 2008.
- [106] Max J Pachl. 3D model facilitated zero-ischemia laparoscopic nephron sparing resection in nephroblastomatosis following the addition of cis-retinoic acid. *Urology*, 138:18–151, 2020.
- [107] Yongsheng Pan, Mingxia Liu, Chunfeng Lian, Tao Zhou, Yong Xia, and Dinggang Shen. Synthesizing missing PET from MRI with cycle-consistent generative adversarial networks for Alzheimer’s disease diagnosis. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 455–463, 2018.
- [108] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035, 2019.
- [109] Francesco Porpiglia, Enrico Checcucci, Daniele Amparore, Federico Piramide, Gabriele Volpi, Stefano Granato, Paolo Verri, Matteo Manfredi, Andrea Bellin, Pietro Piazzola, Riccardo Autorino, Ivano Morra, Cristian Fiori, and Alex Mottrie. Three-dimensional augmented reality robot-assisted partial nephrectomy in case of complex tumours (PADUA ≥ 10): A new intraoperative tool overcoming the ultrasound guidance. *European Urology*, (78(2)), 2020.
- [110] Ahmet K Poyraz, Faith Firdolas, Mehmet R Onur, and Ercan Kocako. Evaluation of left renal vein entrapment using multidetector computed tomography. *Acta radiologica*, 54, 2012.
- [111] R Putz and R Pabst. *Sobotta Atlas of human anatomy: V.2*, 2006.
- [112] G Quero, A Lapergola, L Soler, M Shabaz, A Hostettler, T Collins, J Marescaux, D Mutter, M Diana, and Patrick Pessaux. Virtual and augmented reality in oncologic liver surgery. *Surgical Oncology Clinics of North America*, 28:31–44, 2019.
- [113] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016.
- [114] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2017.
- [115] Blaine Rister, Kaushik Shivakumar, Tomomi Nobashi, and Daniel L Rubin. Data from CT-ORG - The Cancer Imaging Archive, 2019.
- [116] Mateus Riva, Pietro Gori, Florian Yger, and Isabelle Bloch. Is the U-Net directional-relationship aware? In *International Conference on Image Processing (ICIP)*, 2022.

- [117] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, volume LNCS 9351, pages 234–241, 2015.
- [118] Holger R Roth, Amal Farag, Evrim B Turkbey, Le Lu, Jiamin Liu, and Ronald M Summers. Data from Pancreas-CT - The Cancer Imaging Archive, 2016.
- [119] Veit Sandfort, Ke Yan, Perry J Pickhardt, and Ronald M Summers. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Scientific reports*, 9(1):1–9, 2019.
- [120] Edgar Schonfeld, Bernt Schiele, and Anna Khoreva. A U-Net based discriminator for generative adversarial networks. In *IEEE Computer Vision and Pattern Recognition Conference (CVPR)*, pages 8207–8216, 2020.
- [121] Minkyoo Seo, Dongkeun Kim, Kyongmoon Lee, Seunghoon Hong, Jae Seok Bae, Jung Hoon Kim, and Suha Kwak. Neural contrast enhancement of CT image. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3972–3981, 2021.
- [122] Kumar K Shashi, Ted Lee, Sila Kurugol, Harsha Garg, Sunil J Ghelani, Caleb P Nelson, and Jeanne S Chow. Normative values for ureteral diameter in children. *Pediatric Radiology*, 52(8):1492–1499, 2022.
- [123] Santiago Silva, Andre Altmann, Boris Gutman, and Marco Lorenzi. Fed-BioMed: A general open-source frontendframework for federated learning in healthcare. pages 201–210, 2020.
- [124] K Smith, K Clark, W Bennett, T Nolan, J Kirby, M Wolfsberger, J Moulton, B Vendt, and J Freymann. Data from CT-COLONOGRAPHY - The Cancer Imaging Archive, 2015.
- [125] Chongchong Song, Baochun He, Hongyu Chen, Shuangfu Jia, Xiaoxia Chen, and Fucang Jia. Non-contrast CT liver segmentation using CycleGAN data augmentation from contrast enhanced CT. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) iMIMIC Workshop*, pages 122–129, 2020.
- [126] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *International Conference on Computer Vision (ICCV)*, pages 843–852, 2017.
- [127] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15(29), 2015.
- [128] Ahmed Taha, Pechin Lo, Junning Li, and Tao Zhao. Kid-Net: Convolution networks for kidney vessels segmentation from CT-volumes. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2018.
- [129] Zimeng Tan, Jianjiang Feng, and Jie Zhou. SGNet: Structure-aware graph-based network for airway semantic segmentation. In *Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 153–163, 2021.

- [130] Marouane Tilba, Aymen Sekhri, and Aladine Chetouani. Évaluation de la qualité des images médicales basée sur un apprentissage par adaptation au domaine. In *Colloque francophone de Traitement du Signal et des Images (GRETSI)*, 2022.
- [131] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaa El-Nouby, Edouard Grave, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, and Hervé Jégou. ResMLP: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP, 2022.
- [132] Alessio Virzi, Pietro Gori, Cécile Muller, Eva Mille, Quoc Peyrot, Laureline Berteloot, Nathalie Boddaert, Sabine Sarnacki, and Isabelle Bloch. Segmentation of Pelvic Vessels in Pediatric MRI Using a Patch-Based Deep Learning Approach. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) PIPPI Workshop*, pages 97–106, 2018.
- [133] Nicole Wake, James Wysock, Marc Bjurlin, Hersh Chandarana, and William Huang. Pin the tumor on the kidney: An evaluation of how surgeons translate CT and MRI data to 3D models. *Urology*, 131:255–261, 2019.
- [134] Chenglong Wang, Yuichiro Hayashi, Masahiro Oda, Hayato Itoh, Takayuki Kitasaka, Alejandro Frangi, and Kensaku Mori. Tubular structure segmentation using spatial fully connected network with radial distance loss for 3D medical images. In *Medical Image Computing e Computer Assisted Intervent (MICCAI)*, pages 348–356, 2019.
- [135] Chenglong Wang, Masahiro Oda, Yuichiro Hayashi, Yasushi Yoshino, Tokunori Yamamoto, Alejandro Frangi, and Kensaku Mori. Tensor-cut: A tensor-based graph-cut blood vessel segmentation method and its application to renal artery segmentation. *Medical Image Analysis*, 60:101623, 12 2019.
- [136] Di Wang, Yue Pan, Oguz Durumeric, Joseph Reinhardt, Eric Hoffman, Joyce Schroeder, and Gary Christensen. PLOSL: Population learning followed by one shot learning pulmonary image registration using tissue volume preserving and vesselness constraints. *Medical Image Analysis*, 79:102434, 4 2022.
- [137] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 335:34 – 45, 2019.
- [138] Guotai Wang, Maria Zuluaga, Wenqi Li, Rosalind Aughwane, Premal Patel, Michael Aertsen, Tom Doel, Anna David, Jan Deprest, Sebastien Ourselin, and Tom Vercauteren. DeepIGeoS: A deep interactive geodesic framework for medical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1559–1572, 2017.
- [139] Yan Wang, Xu Wei, Fengze Liu, Jieneng Chen, Yuyin Zhou, Wei Shen, Elliot Fishman, and Alan Yuille. Deep distance transform for tubular structure segmentation in CT scans. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3832–3841, 2020.

- [140] Lianne Wellens, Jene Meulstee, Cornelis Ven, C. Scheltinga, Annemieke Littooij, Marry Heuvel-Eibrink, Marta Fiocco, Anne Rios, Thomas Maal, and Marc Wijnen. Comparison of 3-Dimensional and augmented reality kidney models with conventional imaging data in the preoperative assessment of children with wilms tumors. *JAMA Netw Open*, 2(4):e192633, 2019.
- [141] Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet strikes back: An improved training procedure in timm. In *Neural Information Processing Systems (NeurIPS)*, 2021.
- [142] Marko Wilke, V J Schmithorst, and Holland Scott K. Normative pediatric brain data for spatial normalization and segmentation differs from standard adult data. *Magnetic Resonance in Medicine*, 50:749–757, 2003.
- [143] Yongjia Xu and Yongzeng Lai. Derivatives of functions of eigenvalues and eigenvectors for symmetric matrices. *Mathematical Analysis and Applications*, 4441(1):251–274, 2016.
- [144] Ke Yan, Le Lu, and Ronald M Summers. Unsupervised body part regression via spatially self-ordering convolutional neural networks. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1022–1025, 2018.
- [145] Wenjun Yan, Yuanyuan Wang, Shengjia Gu, Lu Huang, Fuhua Yan, Liming Xia, and Qian Tao. The domain shift problem of medical image segmentation and vendor-adaptation by U-Net-GAN. In *Medical Image Computing and Computer Assisted Interventions (MICCAI)*, pages 623–631, 2019.
- [146] Chao Yang, Taehwan Kim, Ruizhe Wang, Hao Peng, and C C Jay Kuo. ESTHER: Extremely Simple Image Translation Through Self-Regularization. In *British Machine Vision Conference (BMVC)*, 2018.
- [147] Heran Yang, Jian Sun, Aaron Carass, Can Zhao, Junghoon Lee, Jerry Prince, and Zongben Xu. Unsupervised MR-to-CT synthesis using structure-constrained CycleGAN. *IEEE Transaction on Medical Imaging*, 39(12):4249–4261, 2020.
- [148] Linlin Yao, Pengbo Jiang, Jon Xue, Yiqiang Zhan, Dijia Wu, Lichi Zhang, Qian Wang, and Feng Shi. Graph convolutional network based point cloud for head and neck vessel labeling. In *Medical Image Computing and Computer Assisted Intervention (MICCAI) MLMI Challenge*, pages 474–483, 2021.
- [149] Xin Yi, Ekta Walia, and Paul Babyn. Generative Adversarial Network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.
- [150] Chenyu You, Guang Li, Yi Zhang, Xiaoliu Zhang, Hongming Shan, Mengzhou Li, Shenghong Ju, Zhen Zhao, Zhuiyang Zhang, Wenxiang Cong, Michael Vannier, Punam Saha, Eric Hoffman, and Ge Wang. CT super-resolution GAN constrained by the identical, residual, and cycle learning ensemble (GAN-CIRCLE). *IEEE Transactions on Medical Imaging*, 39(1):188–203, 2019.
- [151] Qian Yu, Yinghuan Shi, Jinquan Sun, Yang Gao, Yakang Dai, and Jianbing Zhu. Crossbar-Net: A novel convolutional network for kidney tumor segmentation in CT images. *IEEE Transactions on Image Processing*, 28(8):4060–4074, 2019.

- [152] XiaoDong Yuan, Jing Zhang, ChangBin Quan, Yuan Tian, Hong Li, and GuoKun Ao. A simplified whole-organ CT perfusion technique with biphasic acquisition: Preliminary investigation of accuracy and protocol feasibility in kidneys. *Radiology*, 279(1):254–261, 2016.
- [153] B I Yuh and R H Cohan. Different phases of renal enhancement: role in detecting and characterizing renal masses during helical CT. *AJR Am J Roentgenol*, 173(3):747–755, 1999.
- [154] Matthew Zeiler, Dilip Krishnan, Graham Taylor, and Robert Fergus. Deconvolutional networks. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2528–2535, 2010.
- [155] Rongjian Zhao, Buyue Qian, Zhang Xianli, Yang Li, Rong Wei, Yang Liu, and Yinggang Pan. Rethinking Dice loss for medical image segmentation. In *IEEE International Conference on Data Mining (ICDM)*, pages 851–860, 2020.
- [156] Zhongchen Zhao, Huai Chen, and Lisheng Wang. A coarse-to-fine framework for the 2021 kidney and kidney tumor segmentation challenge. In *Medical Image Computing and Computer Assisted Intervention (MICCAI) KiTS Challenge*, pages 53–58, 2021.
- [157] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017.
- [158] Yingying Zhu, Youbao Tang, Yuxing Tang, Daniel C Elton, Sungwon Lee, Perry J Pickhardt, and Ronald M Summers. Cross-domain medical image translation by shared latent Gaussian mixture model. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 379–389, 2020.

Titre: Apprentissage de jumeaux numériques anatomiques en imagerie pédiatrique 3D pour la chirurgie des cancers du rein

Mots clés: intelligence artificielle, Machine learning, applications de santé, big data

Résumé: Les cancers rénaux pédiatriques représentent 9% des cancers pédiatriques avec un taux de survie de 9/10 au prix de la perte d'un rein. La chirurgie d'épargne néphronique (NSS, ablation partielle du rein) est possible si le cancer répond à des critères précis (e.g. le volume et la localisation de la lésion). L'indication de la NSS repose sur l'imagerie préopératoire, en particulier la tomographie informatisée à rayons X (CT). Si l'évaluation de tous les critères sur des images 2D n'est pas toujours facile, les modèles 3D spécifiques au patient offrent une solution prometteuse. La construction de modèles 3D de l'anatomie rénale basés sur la segmentation est développée chez les adultes mais pas chez les enfants. Il existe un besoin de méthodes de traitement d'image dédiées aux patients pédiatriques en raison des spécificités de ces images, comme l'hétérogénéité de la pose et de la taille des structures. De plus, dans les images CT, l'injection d'un agent de contraste est souvent utilisée (ceCT) pour faciliter l'identification de l'interface entre les différentes structures mais cela peut conduire à une hétérogénéité dans le contraste de certaines structures anatomiques, même parmi les patients acquis avec la même procédure.

Le premier objectif de cette thèse est d'effectuer une segmentation des organes/tumeurs à partir d'images ceCT, à partir de laquelle un modèle 3D sera dérivé. Des approches d'apprentissage par transfert (des données adultes aux images enfants) sont proposées. La première question consiste à savoir si de telles méthodes sont réalisables, malgré la différence structurelle évidente entre les ensembles de données. Une deuxième question porte sur la possibilité de remplacer les techniques standard d'augmentation des données par des techniques d'homogénéisation des données utilisant des Spatial Transformer Networks, améliorant ainsi le temps d'apprentissage, la mémoire requise et les performances.

La segmentation de certaines structures anatomiques dans des images

ceCT peut être difficile à cause de la variabilité de la diffusion du produit de contraste. L'utilisation combinée d'images CT sans contraste (CT) et ceCT atténue cette difficulté, mais au prix d'une exposition doublée aux rayonnements. Le remplacement d'une des acquisitions CT par des modèles génératifs permet de maintenir la performance de segmentation, en limitant les doses de rayons X. Un deuxième objectif de cette thèse est de synthétiser des images ceCT à partir de CT et vice-versa, à partir de bases d'apprentissage d'images non appariées, en utilisant une extension des Cycle Generative Adversarial Networks. Des contraintes anatomiques sont introduites en utilisant le score d'un Self-Supervised Body Regressor, améliorant la sélection d'images anatomiquement appariées entre les deux domaines et renforçant la cohérence anatomique.

Un troisième objectif de ce travail est de compléter le modèle 3D d'un patient atteint d'une tumeur rénale en incluant également les artères, les veines et les uretères. Une étude approfondie et une analyse comparative de la littérature sur la segmentation des structures tubulaires anatomiques sont présentées. En outre, nous présentons pour la première fois l'utilisation de la fonction "vesselness" comme fonction de perte pour l'entraînement d'un réseau de segmentation. Nous démontrons que la combinaison de l'information sur les valeurs propres avec les informations structurelles d'autres fonctions de perte permet d'améliorer les performances.

Enfin, nous présentons un outil développé pour utiliser les méthodes proposées dans un cadre clinique réel ainsi qu'une étude clinique visant à évaluer les avantages de l'utilisation de modèles 3D dans la planification préopératoire. L'objectif de cette recherche est de démontrer, par une évaluation rétrospective d'experts, comment les critères du NSS sont plus susceptibles d'être trouvés dans les images 3D que dans les images 2D. Cette étude est toujours en cours.

Title: Learning anatomical digital twins in pediatric 3D imaging for renal cancer surgery

Keywords: artificial intelligence, Machine learning, health applications, big data

Abstract: Pediatric renal cancers account for 9% of pediatric cancers with a 9/10 survival rate at the expense of the loss of a kidney. Nephron-sparing surgery (NSS, partial removal of the kidney) is possible if the cancer meets specific criteria (regarding volume, location and extent of the lesion). Indication for NSS is relying on preoperative imaging, in particular X-ray Computerized Tomography (CT). While assessing all criteria in 2D images is not always easy nor even feasible, 3D patient-specific models offer a promising solution. Building 3D models of the renal tumor anatomy based on segmentation is widely developed in adults but not in children. There is a need of dedicated image processing methods for pediatric patients due to the specificities of the images with respect to adults and to heterogeneity in pose and size of the structures (subjects going from few days of age to 16 years). Moreover, in CT images, injection of contrast agent (contrast-enhanced CT, ceCT) is often used to facilitate the identification of the interface between different tissues and structures but this might lead to heterogeneity in contrast and brightness of some anatomical structures, even among patients of the same medical database (i.e., same acquisition procedure). This can complicate the following analyses, such as segmentation.

The first objective of this thesis is to perform organ/tumor segmentation from abdominal-visceral ceCT images. An individual 3D patient model is then derived. Transfer learning approaches (from adult data to children images) are proposed to improve state-of-the-art performances. The first question we want to answer is if such methods are feasible, despite the obvious structural difference between the datasets, thanks to geometric domain adaptation. A second question is if the standard techniques of data augmentation can be replaced by data homogenization techniques using Spatial Transformer Networks (STN), improving training time, memory requirement and performances.

In order to deal with variability in contrast medium diffusion, a second objective is to perform a cross-domain CT image translation from ceCT to contrast-free CT (CT) and vice-versa, using Cycle Generative Adversarial Network (CycleGAN). In fact, the combined use of ceCT and CT images can improve the segmentation performances on certain anatomical structures in ceCT, but at the cost of a double radiation exposure. To limit the radiation dose, generative models could be used to synthesize one modality, instead of acquiring it. We present an extension of CycleGAN to generate such images, from unpaired databases. Anatomical constraints are introduced by automatically selecting the region of interest and by using the score of a Self-Supervised Body Regressor, improving the selection of anatomically-paired images between the two domains (CT and ceCT) and enforcing anatomical consistency.

A third objective of this work is to complete the 3D model of patient affected by renal tumor including also arteries, veins and collecting system (i.e. ureters). An extensive study and benchmarking of the literature on anatomic tubular structure segmentation is presented. Modifications to state-of-the-art methods for our specific application are also proposed. Moreover, we present for the first time the use of the so-called vesselness function as loss function for training a segmentation network. We demonstrate that combining eigenvalue information with structural and voxel-wise information of other loss functions results in an improvement in performance.

Eventually, a tool developed for using the proposed methods in a real clinical setting is shown as well as a clinical study to further evaluate the benefits of using 3D models in pre-operative planning. The intent of this research is to demonstrate through a retrospective evaluation of experts how criteria for NSS are more likely to be found in 3D compared to 2D images. This study is still ongoing.