Article

Multi-Modal U-net for Segmenting Gross Tumor Volume in Lungs during Radiotherapy

Dhruv Jain [1,*], Romain Modzelewski [3,2], Romain Herault [1], Clément Chatelain [1] and Sébastien Thureau [4,2]

[1] INSA Rouen Normandie, Univ Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen, France
[2] Univ Rouen Normandie, INSA Rouen Normandie, Université Le Havre Normandie, Normandie Univ, LITIS UR 4108, F-76000 Rouen, France
[3] Nuclear Medicine Department, Henri Becquerel Cancer Center, Rouen, France
[4] Radiotherapy Department, Henri Becquerel Cancer Center, Rouen, France
* Corresponding Author, Email-dhruv.jain@insa-rouen.fr

Abstract: In this work we introduce an end-to-end multi-modal neural network to segment the Gross Tumor Volume (GTV) from 3D-CBCT's during radiotherapy. We improve the tumor segmentation by using a U-net which takes additional information such as the tumor mask generated at the planning phase along with the CBCT volume. The mask contains relevant information about the tumor's location and can guide the model to use this knowledge appropriately to give a better prediction. This technique could become an alternative to produce segmentation masks of GTV in CBCT automatically during radiotherapy as in the traditional RT-pipeline, they are not segmented. We have evaluated our model on a dataset of 82 patients who have undergone radiotherapy. We compare the results of registered target volumes from planning CT as mask seed with 2 different types of multi-modal architectures. Our model shows a DSC of 0.706±0.002 with Late Fusion and 0.702±0.015 with Early Fusion using the GTV Mask. The performance of the two models on this mask is similar, so we perform further experiments with different types of masks which suggest that Late Fusion model produces a better segmentation of the tumor than the Early Fusion model. We also provide an ablation study consisting of a single modality U-net and a metric based on the Planning CT mask registration. It indicates a clear advantage of using our model to produce segmentation for this type of imaging.

Keywords: Deep Learning; Semantic Segmentation; Radiotherapy; Multimodality

1. Introduction

Lung cancer is the leading cause of cancer related deaths which contributes upto 18% of the total deaths which is estimated to be around 1.8 million [1]. Around 30-60% of patients having lung cancer go through radiotherapy during their treatment [2]. Increased dosage of X-rays on the cancer cells can break and destabilise its DNA (healthy cells are also disturbed but have better recovery mechanism) and hence can stop them from growing and might be able to reduce their volumes. [3]. There are multiple phases in classical Radiotherapy pipeline such as the simulation, planning, and delivery [4]. In the planning phase, a CT is acquired and annotated by experts to provide the accurate delineation of the target and the organs at risk volume. Subsequently a plan is formulated and the shape of the beams is calculated by inverse methods which could provide the prescribed dose distribution. In the delivery phase, radiation is applied to the targeted region, while a pre-treatment CBCT is acquired for accurate positioning of the patient's bone anatomy with the treatment couch. Figure 1 shows the classical radiotherapy pipeline. In this framework, CBCT delineations, if existing, are never used. With the rise of new technology and approaches such as Adaptive Radiotherapy(ART), it might be useful to segment CBCT's. ART is used to manage the change in the anatomical/functional structures which could be caused due to various factors such as weight loss in the patient, shrinkage in the targeted

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

area or inflammation in the patient's body. Re-planning is a procedure in the course of ART which selects the appropriate patient who could benefit by adopting a new plan and the correct stage at which there is significant change in the anatomy. Since CBCT's are the most recent images of the patient and they can guide us with the evolution of the tumor from the planning CT, it is important to acquire its segmentation for further analysis of the growth of the tumor.

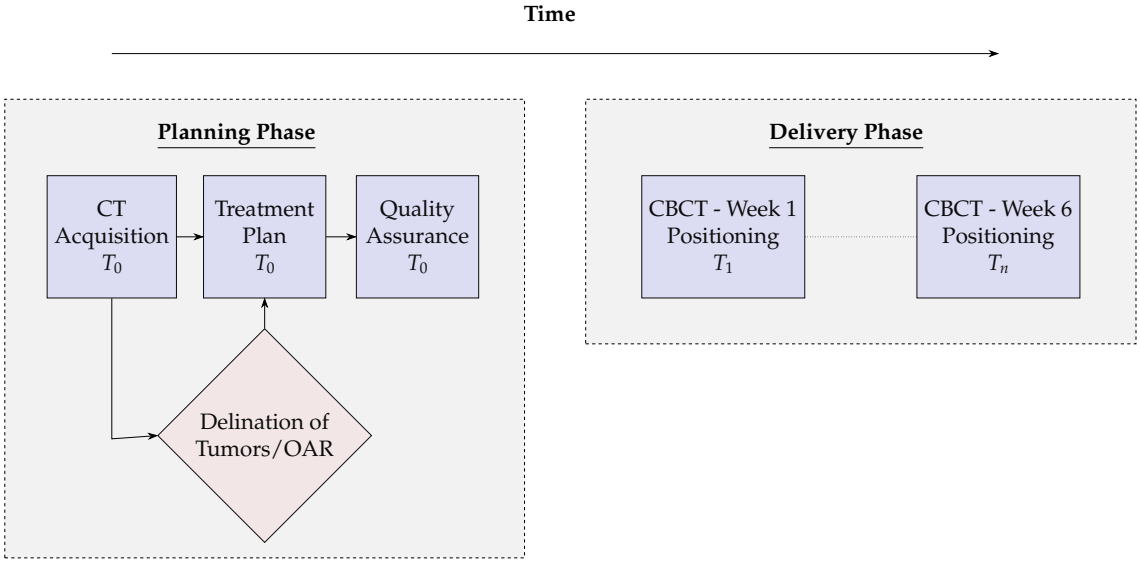

Figure 1. Traditional Radiotherapy showing the Planning and the Delivery of the Treatment

Automatic Delineation of the Tumor Volumes(TV) in CBCT can be useful to make an automated decision about re-planning, without the extra burden on the medical staff. Deformable Image Registration(DIR) is used for this purpose in which the Planning CT is deformed to resemble the target image and then the same transformations are applied on the Planning CT-masks which could give a close approximate of the target mask. Although the quality of the segmentation highly depends on the type of registration, and the Planning CT's closeness to the daily image [5], moreover this method does not produce the perfect segmentation of the CBCT and manual intervention is required to correct the contours. CT-CBCT registration creates further problems as there are inconsistent intensities and so an intensity correction of a voxel in CBCT is required [6]. Furthermore issues such as tumor shrinkage from the time the CT had been taken, inter and intra observer segmentation variability of the CT-mask needs to be taken care. Deep Learning architectures have been getting more and more popular in the medical field. Convolutional Neural Networks provide an elegant and efficient way to learn both a feature extraction model and a decision model in an end-to-end manner [7]. With the use of gpu programming and extracting parallelism in training, we are able to train deeper networks [8] and architectures which could be used for a wide range of problems such as classification, semantic segmentation, object detection, domain adaptation and generative models. U-net architecture is very popular in medical imaging for semantic segmentation [9]. It is a fully convolutional network where a pixel-wise loss is calculated between the ground-truth and the prediction. The main contribution of U-nets and the reason for its popularity are its long skip connections. ie. The down-sampled feature maps are concatenated with the up-sampled feature maps. The skip connection provides another way for the gradient to flow and it helps the network to learn the lower level features which might get corrupted in case of auto-encoders. These networks can be used for 3D-Data [10] as well, where we just replace the 2D-Convolution and 2D-Max-pooling with a 3D-Convolution and 3D-Max-pooling respectively.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

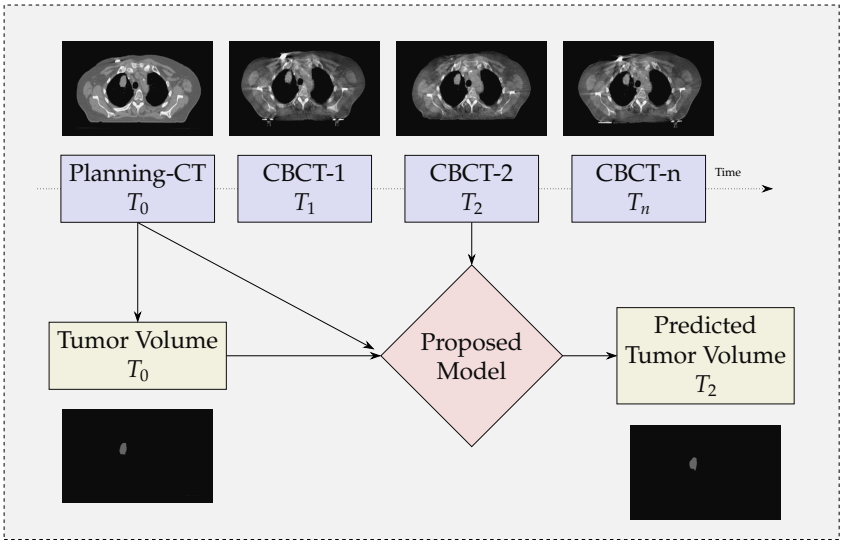

Figure 2. Overview of the proposed pipeline for tumor volume segmentation

1.1. Proposition

Segmentation in CBCT is difficult due to soft tissue contrast and presence of scatter artifacts. Due to these reasons it is difficult to perform a DIR between CT and CBCT. Furthermore, it can be seen from our experiments that using a single modality Unet might not be sufficient enough to delineate GTV in CBCT. Although it is important to delineate CBCT for various tasks, there is no clear efficient method which is standardised or is being practiced. In this work, our contributions are as follows -

- We propose a multi-modal neural-network which uses CBCT and a registered CT-Mask produced during Planning phase(as shown in Figure 2) to train an end-to-end 3D U-net to automatically delineate the Gross Tumor Volume in the CBCT. It produces reasonably accurate contours of GTV during Radiotherapy.
- We provide a comparison between two types of fusion - Early Fusion and Late Fusion by using different types of imprecise CT-masks. This helps to take a better decision in choosing the architecture.

Though the imprecise CT-masks should not be used to segment the tumor volumes clinically, we used it to get a better comparison of both the architectures of the U-net. We have organised the rest of the paper as follows. Section 2 describes some of the significant works related to our study, Section 3 is devoted to provide details of the registration and our network. Section 4 describes the information related to our experiments and the analysis of our results. We show some illustrations and discuss the performance of our models in Section 5, and finally, we conclude in Section 6.

2. Related Works

Generally, tumors are difficult to segment as they might be found in regions with low contrast and hence it is more difficult to have accurate boundaries. Ge *et al.* [11] introduced the Multi-input dilated (MD) U-net to segment bladder tumor. They mentioned that the traditional U-net down-samples the original features to learn global features, but it ends up in corrupting the local features of small sized objects. They replaced the max-pooling operation of down-sampling with dilated convolution which increases the receptive field. Furthermore they used multiple scaled inputs at different levels so that the context information could be improved. Wang *et al.* [12] proposed a network i.e. A-net for the semantic segmentation of tumors to be used in Adaptive Radiotherapy(ART). Their model used a deep learning network with patches of 3x3 cm as the input. This helped their model with low data size. Patches as input gives more consideration to the local level features rather than global features. Though they used the initial weeks MRI volumes in the training

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

set and the last MRI of the same patients in the testing set. As medical data is always limited in perspective of annotation or data size, use of multiple modalities is getting more popular. Wang *et al.* [12] used CT and MRI to create a 2 stage network which uses Cycle-Gan to learn important features of the tumor. In the first stage, they have a cycle consistency loss between the two domains, while they also introduced another structure loss for the tumor which takes into account the shape and size of the tumor generated by the gan. In the second stage, the pseudo MRI images are collected together with the few available expert-annotated MRI scans to train the network. Li *et al.* [13] have introduced a multi-modal network which uses CT and PET images to segment the tumor. PET images with F-FDG(F-fluorodeoxy-glucose), helps to show a clearer contrast at the tumor boundaries. Since these type of images have low spatial resolution, so a fusion of PET and CT is an interesting approach. They generate a probability map of the tumor from CT using a FCN. Then this map is fused with the intensity values of PET via a fuzzy variational model. Zhao *et al.* [14] also used CT-PET images for segmentation. They introduced two networks . First, a multitask network to extract the features maps from CT and PET images separately. Then, they used another network comprising of cascaded convolution operations which gave the segmentation map. Jin *et al.* [15] introduced the DeepTarget network which could delineate GTV and CTV in CT guided with PET. They used two stream 3D fusion PSNN network based on Unet and PHNN [16]. They first carry out a deformable registration between the CT and the PET and segment GTV and organs at risk to use in the final network for CTV delineation. Wang *et al.* [17] used multi-view fusion segmentation for GTV segmentation of brain glioma on CT images. They used an encoder-decoder architecture similar to a U-net which used 3(current, previous and next) 2D-CT images whose features are fused at the decoder ie:Dense-Decoder. They mentioned that this type of input covered more spatial region than 2D CNN while it had less parameters than 3D CNN. Ma *et al.* [18] proposed a registration-guided deep learning architecture that used CBCT images and registered CT-Masks to delineate Organs at Risk. They used two different types of registration on the CT masks i.e. Rigid and Deformable Registration. They show that the deformable registration performs better than the rigid registration. Segmentation of Cone Beam CT is difficult due to lower soft tissue contrast and generation of artifacts. Fu *et al.* [19] used a cross modality attention pyramid network to automatically segment bladder, prostate, rectum, and left/right femoral heads in CBCT. This network consisted of 2 U-nets which took one of the inputs i.e. CBCT or a synthetic MRI. The loss used for training is a combined loss of the 2 Unet networks and also a loss from the late fusion of the features in the 2 decoders via an attention gate.

$$L = L_{CBCT} + L_{sMRI} + L_{LateFusion} \tag{1}$$

The synthetic MRI is made by training a CycleGan which learns the translation between CBCT and MRI [20]. For this purpose, they performed a rigid registration between the two images. They also mention that errors in the registration could deteriorate the performance of the segmentation by the network. Jia *et al.* [21] used a CycleGan to translate CT (with contours) to a synthetic CBCT (with no contours) and used domain adaptation with adversarial feature learning to train the CBCT segmentation network without any CBCT annotations. They observed that with adversarial learning, the network produced a higher DSC in comparison to the network which used sCBCT directly from the Cycle-Gan.

$$L = L_{adv} + \lambda_{seg} L_{seg} \tag{2}$$

They trained the domain discriminator first until a threshold and then started the training of the CBCT, s-CBCT segmentation network which used the sCBCT contours for calculating the dice loss. Brion *et al.* [22] also used an adversarial network for unsupervised domain adaptation between annotated CT's and non-annotated CBCT's. They used a

 doi:10.20944/preprints202304.0129.v1

3D-Unet which were trained to segment CT images. Along with that they added a gradient reversal layer(GRL) at the decoder which reduced the domain shift between the CT and the CBCT. GRL is a custom layer where the gradients are changed and hard-coded. They also introduced different strategies for intensity based data augmentation. It improves generalization of CT models to use CBCT data without explicitly training with its contours. We realised that there is a need for a segmentation network which uses the data produced during the Planning Phase to delineate GTV in CBCT. Generally, multiple-modality for automatic segmentation would require extra data to be generated which could be considered as burdensome as it would require an additional(MRI or PET) imaging modality. Since in the planning phase we manually annotate the tumor, this information is important and should be used for further delineation of the GTV in the CBCT. Though our approach is similar to Lin Ma et al in terms of the input, but we differ from their approach as we use the simplest form of registration i.e. Translation. We further perform analysis of different fusion strategies using different types of inaccuracies in the CT-Mask. As mentioned in [18], the OAR and the tumor volume are required to be delineated. As they segment the organs at risk, we go forward to segment the GTV using this approach.

3. Materials and Methods

3.1. GTV Seed for Localisation of Tumor

We propose an end-to-end 3D-Unet which uses the CBCT volume for segmentation of Gross Tumor Volume(GTV) in patients undergoing Radiotherapy. GTV segmentation is a difficult task to train any deep learning network, as tumors can be of different shape/size and generally can be formed in low contrast regions [5]. Hence it may be difficult to clearly identify them. Due to this reason we add additional data to guide the model towards the spatial location of the tumor. To help in the localization we use registered CT-masks of the GTV so as to identify the region close enough to the tumor where the network should focus in delineating the GTV accurately. As these masks are the closest possible approximation of the CBCT tumor, they can provide useful information to the network. Though there should be a balance of information, as too much dependence on the mask might affect the overall performance of the model. One of the reasons of using such a method is that CBCT's are only used for image guidance and are rarely delineated manually, and so there exists a problem of limited data [18]. This type of model helps in solving this limitation as it guides the network towards the GTV. This method could help in saving time and resources as we can avoid using difficult algorithms such as atlas, DIR which are dependent on a specialised type of registration. Figure 3 describes our proposed framework for segmenting the gross tumor volume. Here T_0 represents the temporal checkpoint where the Planning CT was acquired, while T_n represents the same for the corresponding CBCT. The CT and CBCT were acquired on different days and so the position of the patient and the tumor might be different. Registration is a common technique used in medical imaging for comparing two images/volumes. It is used to transform different data to the same coordinate system [23]. Different types of registration change the images in different ways. We use a translation registration, so as to align the 3D volume of the CT to the 3D volume of the CBCT. The same transformations are then applied on the CT-masks. This process needs to be examined for each of the patient. As an incorrect registration can produce errors which gets propagated to the segmentation network. We used the Plastimatch library for this task. Entire volumes of different sizes (CT and CBCT) were registered with each other. This operation is very essential for our model as only after the registration, we can obtain the appropriate location of the tumor in CT with respect to the CBCT. We performed registration for each of the tumor volumes ie: GTV, CTV, PTV.

3.2. Network Architecture

We have used multimodal 3D U-net which takes input as CBCT and the registered CT Mask. 3D U-net [10] is one of the most popular networks for 3D Image Semantic Segmentation. We used 6 blocks (Convolutions, Batch Norm and Relu) in the encoder and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

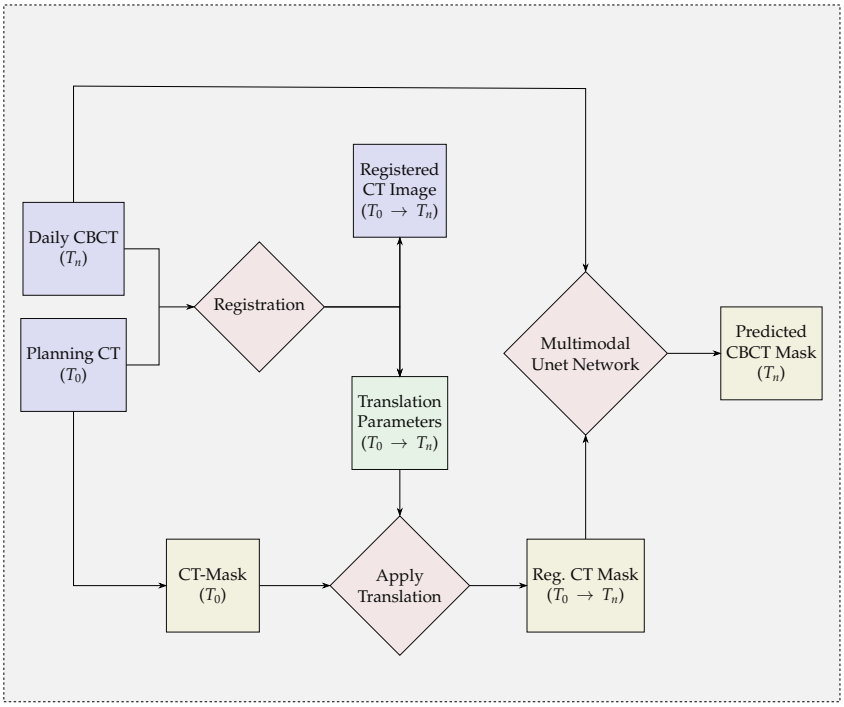

Figure 3. Proposed Framework for Multimodal Segmentation of the Gross Tumor Volume

the decoder part of the U-net, and a bottleneck layer with 2 identical blocks, also we used 3d max-pooling layers to reduce the dimensions and strided convolutions for up-sampling and to get back to the same shape as the segmentation map. We use a 1×1 Conv3d layer after the decoder layer, to change the channel output to 1. There is a sigmoid layer at the end. We calculate the loss between a discrete ground truth and a continuous sigmoid output. We have compared two architectures, i.e. early fusion and late fusion. Figure 4 and Figure 5 shows the two fusion architectures respectively. In early fusion network the information is fused from the first convolution itself while in the late fusion network it takes place at the bottleneck layer. We used these two networks to understand the performance when the registered CT mask is added/removed from the skip connections.

3.3. Evaluation Metric and Loss

We used the Dice loss to train the neural networks which is given by the Equation 3.

$$DiceLoss = 1 - DSC \tag{3}$$

We use a collection of metrics to evaluate our models. Since Recall or Sensitivity penalize errors in smaller segments [24] more than in bigger segments. It is considered as a good measure to check the performance of smaller tumors. Hence, we used the Dice Coefficient(DSC), Recall, Precision and Volume Similarity as a metric to compare different models.

$$Recall = \frac{TP}{TP + FN} \quad \text{and} \quad Precision = \frac{TP}{TP + FP} \tag{4}$$

$$VS = 1 - \frac{||S_t| - |S_g||}{|S_t| + |S_g|} \tag{5}$$

where S_t and S_g are the volumes of the segments we need to compare.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

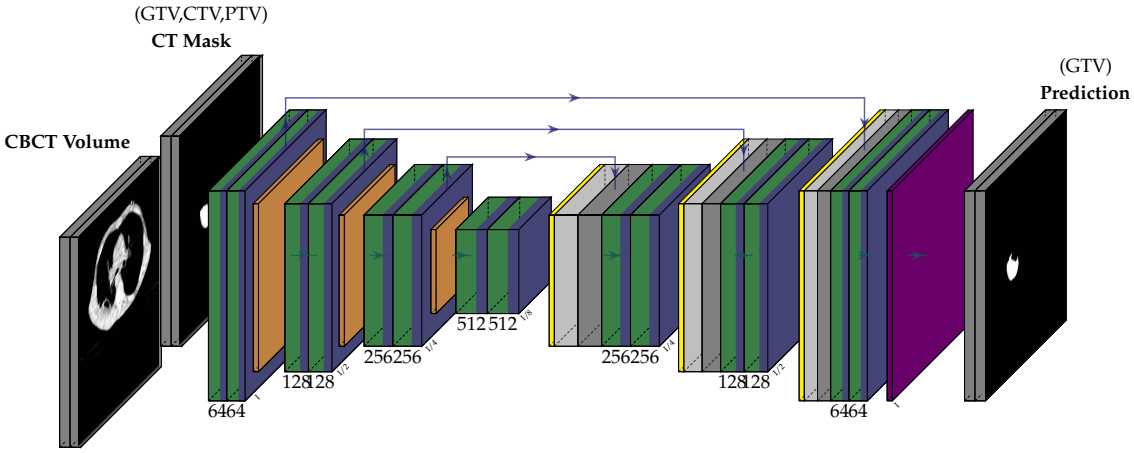

Figure 4. Early Fusion Multimodal 3D-Unet

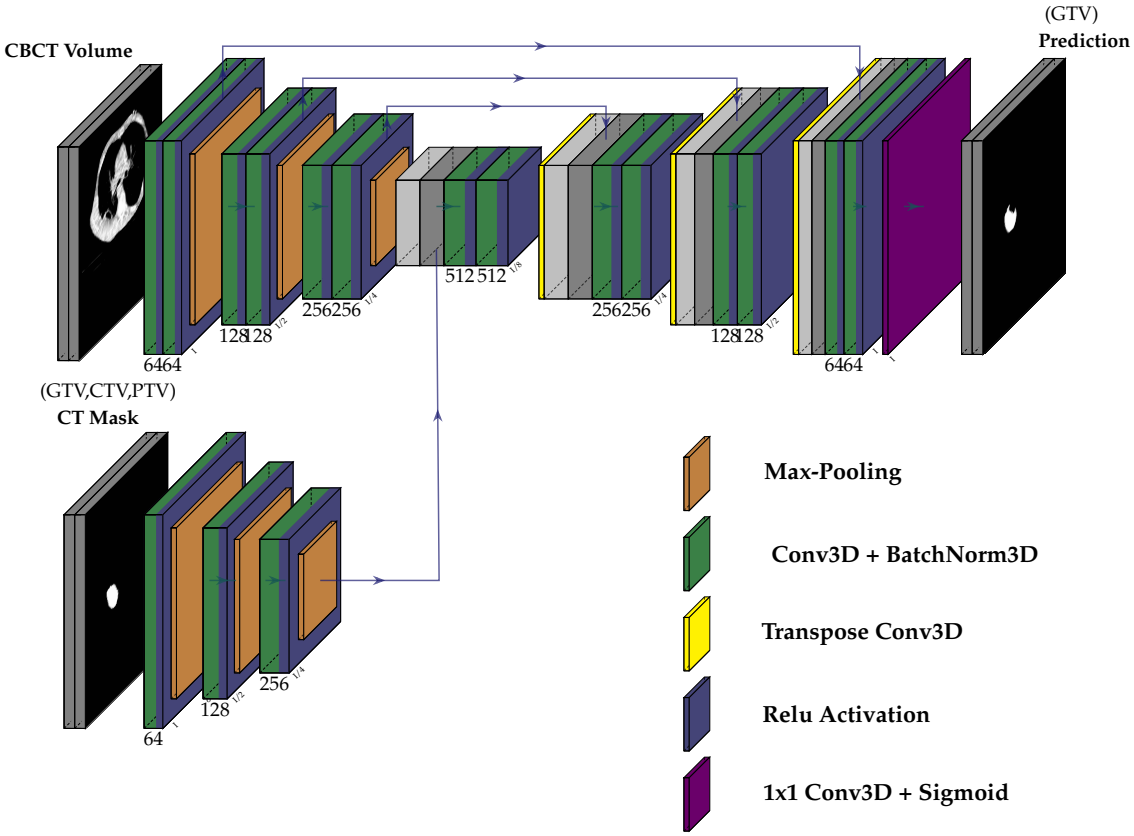

Figure 5. Late Fusion Multimodal 3D-Unet

4. Experiments and Results

4.1. Dataset

The dataset is a private dataset which went through an anonymization process. The patients were being treated for Non-Small Cell Lung Cancer(NSCLC). Eighty-Two patients who went through radiotherapy with non-operated NSCLC were selected. They received 60-70 Gy RT. The procedure for the imaging was 3D Free Breathing acquisition while injecting with an iodine contrast. SIEMENS CT was used for the acquisition of the planning CT and a VARIAN CBCT was used for onboard imagery during radiotherapy. For each patient there were around 6-7 CBCT generated during the radiotherapy process and 1 CT

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

which was generated during the planning process. Each of these CBCT were registered with the one previous to it, while the first one was registered with the CT. The observer pasted the GTV_{n-1} on $CBCT_n$ and performing a threshold[-400, +175 HU] to exclude the healthy tissues and to delineate GTV_n. Finally, this GTV was visually reviewed and manually adapted in case of apparent anatomical changes. Each of the slices were reconstructed to a 512x512 resolution.

4.2. Data Preprocessing

After the registration we resize the CBCT volume to [200,200,64] along with the CT/CBCT masks. The intensities for the CBCT volume is rescaled to [0,1] for faster training. We didn't use any Data Augmentation in our method. Each CBCT is considered as a different instance even if it is from the same patient. We chose 61 randomly selected patients for the training set, 14 for the Validation Set and the remaining 7 in the Test set, while constraining each patient to be found in exactly one of the sets.

4.3. Training

We use the Pytorch framework [25] to construct the neural network. The inputs are a CBCT volume and a Registered CT-Mask volume to the U-net with each of them having the shape of [200,200,64]. We use the Adam optimizer [26] with beta1 = 0.9 and beta2 = 0.999 and a learning rate of 0.0003 with a batch size of 2. It seems that 3D-BatchNorm is important for the model for accurate delineation. We run each model for 100 epochs and then we evaluate the performance of our network by using the best model which had the lowest dice loss in the validation set.

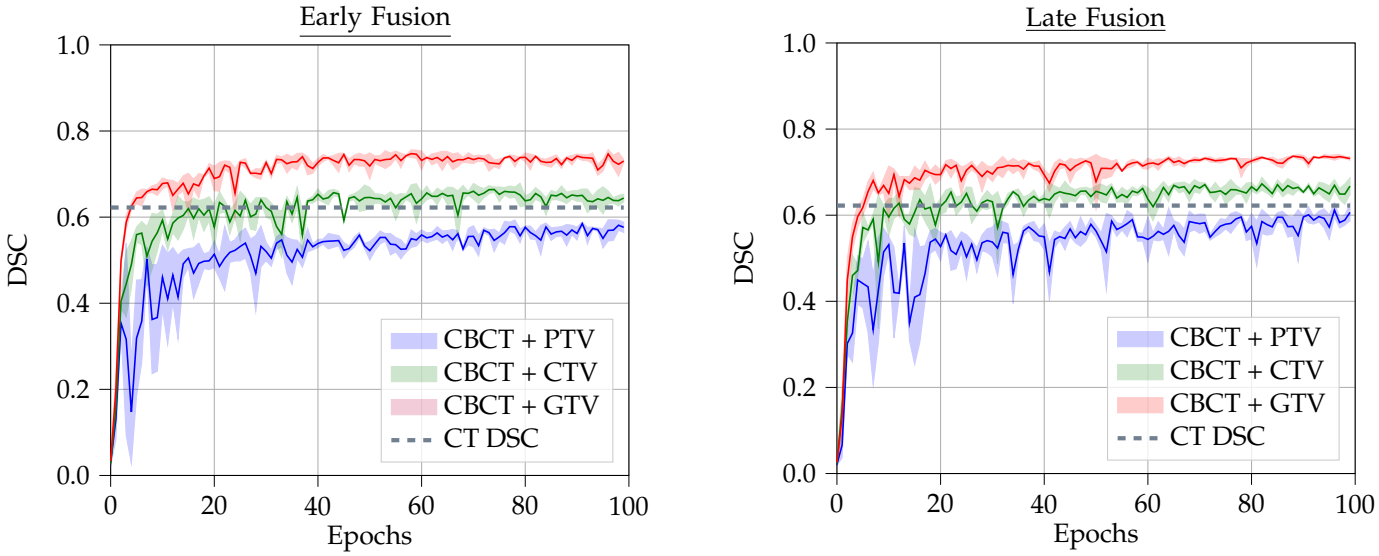

Figure 6. DSC Plot of Validation Set in Early Fusion and Late Fusion

4.4. Results

We show a comparison between 2 different Multimodal U-nets which uses different kinds of CT-Masks. This helps us to compare the different models and it gives us an idea of the better architecture in context of the tumor representation with the help of imprecise masks. In the Early Fusion(EF) network, both the inputs are fused together from the 1st Convolution. The features extracted from these inputs go through the network and the skip connections. While in the Late Fusion(LF) network, the fusion takes place in the bottleneck layer of the U-net, and so the features extracted from the CT-Mask are not present in the skip connections. As the GTV of CT is a close approximation to the ground-truth of the CBCT, it is essential to verify if the CT-Mask provides only the localization information

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

and so it is important to test the training with more imprecise masks. So we compare the different types of inputs mentioned below -

- CBCT + CT-GTV Mask
- CBCT + CT-CTV Mask
- CBCT + CT-PTV Mask

GTV, CTV and PTV masks are delineated manually by radiologists. The target i.e. GTV is the same for all the 3 input masks. CTV and PTV Masks would add imprecision to the GTV contours and so it would be useful to compare these models using these masks. Figure ?? shows the comparison of DSC of the validation set between the two models during the training. The CT DSC in the figure can be considered as a starting point which exhibits the dice coefficient between the GTV CT-Mask and the GTV Ground-truth. It can be considered as a baseline from where we improve our model. The difference between this value and the CBCT+GTV, indicates the improvement in the prediction over the CT-Mask. Though both the plots in the figure seem quite similar, there is a slight deviation in the training of the CBCT + CT-PTV mask which causes a major difference in the performance of the test-set. In this case, LF converges faster than EF, and also the gap between the 3 masks is lower in LF than in EF with similar performance.

CBCT	Fusion	Tumor Mask	DSC	VS	Recall	Precision
Yes	EF	GTV	0.702±0.015	0.837±0.037	**0.845±0.007**	**0.853±0.010**
Yes	LF	GTV	**0.706±0.002**	**0.859±0.018**	0.824±0.003	0.818±0.006
Yes	EF	CTV	0.680±0.017	0.839±0.022	0.804±0.013	0.735±0.057
Yes	LF	CTV	**0.708±0.028**	**0.850±0.052**	**0.822±0.011**	**0.740±0.022**
Yes	EF	PTV	0.460±0.016	0.667±0.113	**0.788±0.019**	0.465±0.089
Yes	LF	PTV	**0.665±0.012**	**0.860±0.028**	0.787±0.009	**0.686±0.033**
Yes	NA	NA	0.425±0.025	0.574±0.020	0.608±0.037	0.266±0.041
No	NA	GTV	0.577			
No	NA	CTV	0.378			
No	NA	PTV	0.189			

Table 1. Comparison of the models with different types of Masks as Input - Rows 1-6 show the full model, Last four rows show the ablation study - (row 7) with single modality U-Net using CBCT as input and (rows 8-10) only the TV registration

Table 1 shows the performance of both the models on each of the type of masks. We can observe that the LF outperforms EF in all types of the masks. Even though the DSC in CBCT with GTV mask for EF is close to the DSC of LF, it is interesting to see that the CBCT with PTV mask of LF performs fairly better. Furthermore, the Volume Similarity(VS) is always higher in LF than in EF, while using the same CT-Mask. This would suggest that LF's volume is closer to the ground-truth's volume. Since LF does not use the mask in the skip connection, it can be reasoned that this model is less dependent on the mask and so it is having a better representation of the tumor.

Ablation study is displayed in the last 4 rows. On the 4th last row, we gave the current CBCT information only to the segmentation process and did not use the registered TV from planning CT. It shows a clear detrimental gap in performance. The 3 last rows correspond to the opposite: It shows only the registration from planning CT to CBCT without using a U-Net network. These lines represents the DSC between the different registered CT masks and the GTV ground-truth. It can be seen that the accuracy of the registered masks is improved by these models and they can represent the tumor volume in a better way.

5. Illustrations

The tumors which are attached to the lung wall nodules is generally difficult to segment as they are in low contrast regions. Though operations such as threshold and reshaping the tumor can be appropriately done by the network, problem arises when the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

tumor lie is in these difficult regions. Hence the network uses the CT-Mask for its accurate delineation, but too much dependency on the CT-Mask may prevent the network to learn accurate features of the tumor. Therefore, there should be a critical balance to learn the features between the CT-Mask and the CBCT Volume by the network. Figure 7 shows a comparison of the test images by using different masks on the two different types of fusion. We display the middle slice of the tumor where the density of tumor pixels is generally high.

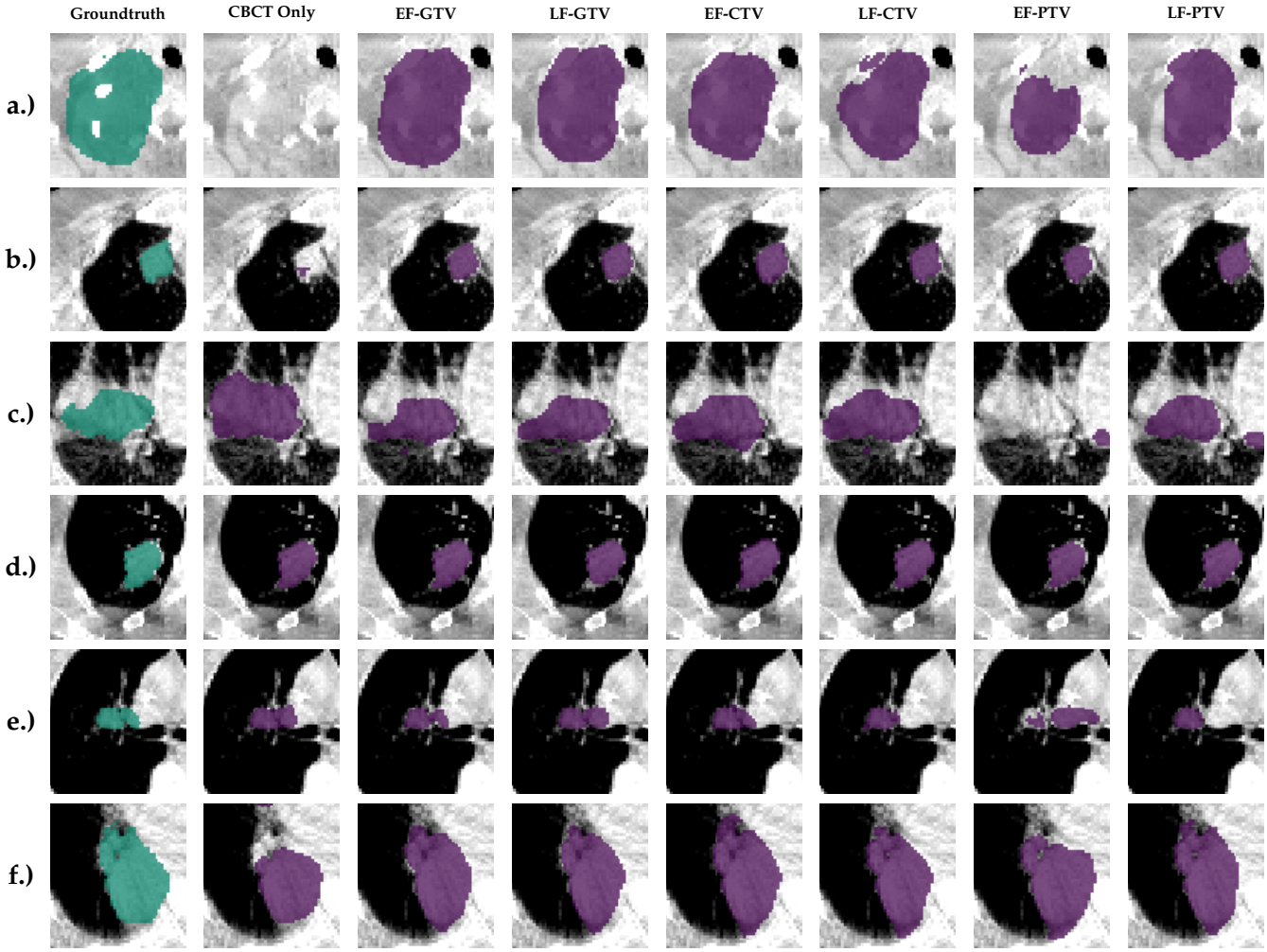

Figure 7. Comparison of the Prediction in Test Set

In single modality i.e. CBCT Only, we can observe that the model learns about the tumor location, but in most of the low contrast regions it fails to identify the tumor(As can be seen in Figure 7(b.) and (c.)). Due to this reason we require another modality as an input to the network. Moreover in these instances the physicians needed an additional PET image to correctly identify the functional part of the tumor and delineate them as shown in the Figure 8.

Hence, the registered CT Mask helps the model to identify the tumor location in low contrast. Furthermore, we noticed that in the EF-PTV model, the network gets too much dependent on the imprecise CT-Mask, and hence affects its performance. The late fusion shows improved performance as the CT-Mask is not included in the skip connections. Also we observed that none of the models could accurately remove the bones which could be seen in Figure 7(a.), as they are not found in the CT-Mask which indicates that the performance of the model could be improved by being more dependent on the features from the CBCT Volume than the CT-Mask. In addition, this patient has the tumor found on

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 7 April 2023 doi:10.20944/preprints202304.0129.v1

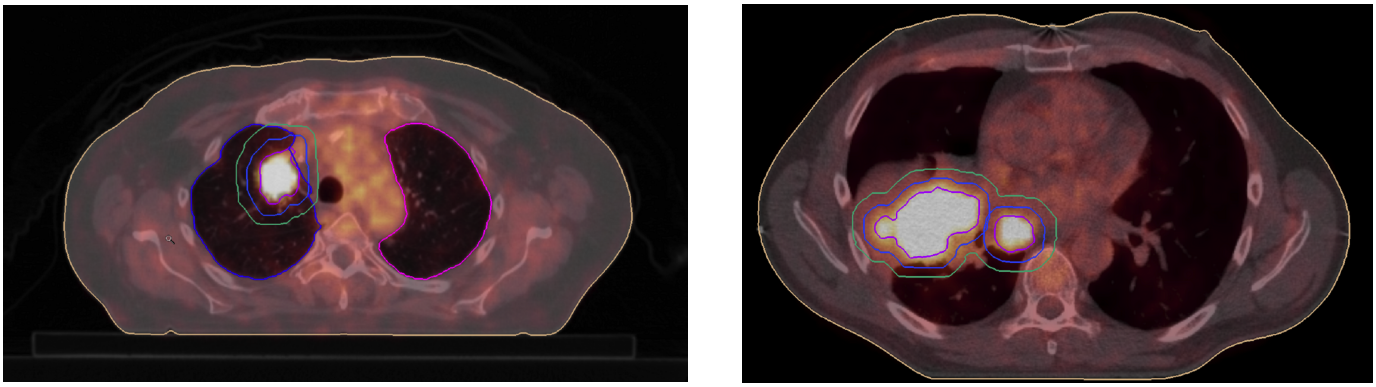

Figure 8. PET Scan Images used for the Delineation of the Tumors for patients having tumors in low contrast regions.

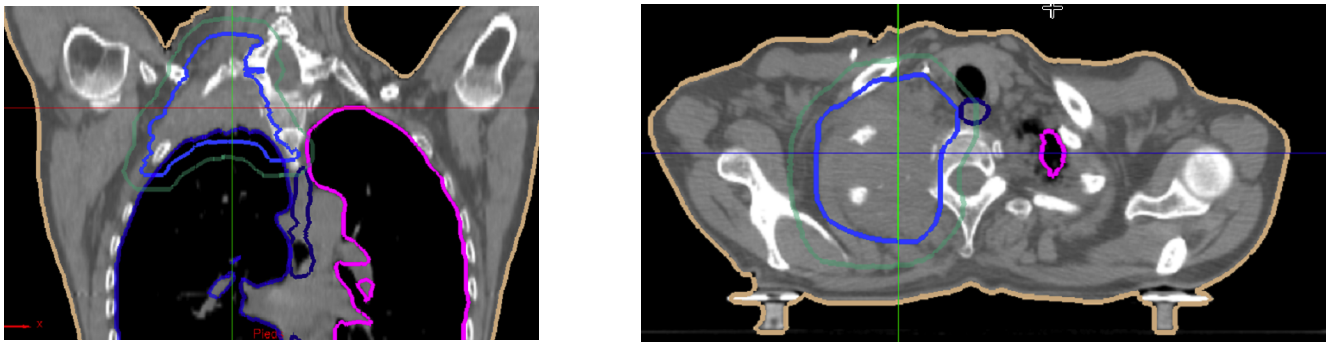

Figure 9. CT Image(Coronal and Axial) for the patient having tumor above the lungs.

the location above the lungs and so it adds complexity for the model to identify it. The CT Scan image for the tumor for this patient is shown in Figure 9.

6. Conclusion

We put forward an end-to-end multi-modal network based on the popular 3D U-net for the segmentation of tumors during Radiotherapy with the use of simple minimal registration i.e. Translation. This model can be an alternative to the popular atlas method for automatic segmentation which is heavily dependent on the performance of the deformable registration. We compared different types of CT Masks and evaluated two types of fusion techniques between the inputs. In our analysis we found that Late Fusion had a better performance of segmenting the tumor that the Early Fusion Model. In future, we might be able to further improve the performance by using a new loss function which penalises the dependency of the model on the CT-Mask, or even by a different type of fusion technique. It might be a good idea to have a bigger dataset to help in generalization. We are further planning to use this type of model in multi-task learning for regression/classification tasks along with the segmentation.

7. Acknowledgement

The project was supported by the MINMACS Région Normandie excellence label and "ANR-20-LCV1-0009"-ANR-LabCom L-Lisa. We also thank CRIANN [1] for their computational resources for this project.

References

1. Sung, H.; Ferlay, J.; Siegel, R.L.; Laversanne, M.; Soerjomataram, I.; Jemal, A.; Bray, F. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for

[1] Centre des Ressources Informatiques et Applications Numérique de Normandie, France

36 Cancers in 185 Countries. *CA: A Cancer Journal for Clinicians* **2021**, *71*, 209–249. https://doi.org/10.3322/caac.21660.
2. Dong, X.; Lei, Y.; Wang, T.; Thomas, M.; Tang, L.; Curran, W.J.; Liu, T.; Yang, X. Automatic multiorgan segmentation in thorax CT images using U-net-GAN. *Medical Physics* **2019**, *46*, 2157–2168. https://doi.org/10.1002/mp.13458.
3. Baskar, R.; Lee, K.A.; Yeo, R.; Yeoh, K.W. Cancer and radiation therapy: Current advances and future directions. *International Journal of Medical Sciences* **2012**, *9*, 193–199. https://doi.org/10.7150/ijms.3635.
4. Lecchi, M.; Fossati, P.; Elisei, F.; Orecchia, R.; Lucignani, G. Current concepts on imaging in radiotherapy. *European Journal of Nuclear Medicine and Molecular Imaging* **2008**, *35*, 821–837. https://doi.org/10.1007/s00259-007-0631-y.
5. Liu, X.; Li, K.W.; Yang, R.; Geng, L.S. Review of Deep Learning Based Automatic Segmentation for Lung Cancer Radiotherapy. *Frontiers in Oncology* **2021**, *11*. https://doi.org/10.3389/fonc.2021.717039.
6. Zhen, X.; Gu, X.; Yan, H.; Zhou, L.; Jia, X.; Jiang, S.B. CT to cone-beam CT deformable registration with simultaneous intensity correction. *Physics in Medicine and Biology* **2012**, *57*, 6807–6826. https://doi.org/10.1088/0031-9155/57/21/6807.
7. Suzuki, K. Overview of deep learning in medical imaging. *Radiological Physics and Technology* **(2017)**, *10*, 257–273. https://doi.org/10.1007/s12194-017-0406-5.
8. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Neural Information Processing Systems* **2012**. https://doi.org/10.1145/3065386.
9. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Springer International Publishing* **2015**. https://doi.org/10.1007/978-3-319-24574-4_28.
10. Özgün Çiçek.; Abdulkadir, A.; Lienkamp, S.S.; Brox, T.; Ronneberger, O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. *Medical Image Computing and Computer-Assisted Intervention* **2016**. https://doi.org/10.1007/978-3-319-46723-8_49.
11. Ge, R.; Cai, H.; Yuan, X.; Qin, F.; Huang, Y.; Wang, P.; Lyu, L. MD-UNET: Multi-input dilated U-shape neural network for segmentation of bladder cancer. *Computational Biology and Chemistry* **2021**, *93*. https://doi.org/10.1016/j.compbiolchem.2021.107510.
12. Wang, C.; Tyagi, N.; Rimner, A.; Hu, Y.C.; Veeraraghavan, H.; Li, G.; Hunt, M.; Mageras, G.; Zhang, P. Segmenting lung tumors on longitudinal imaging studies via a patient-specific adaptive convolutional neural network. *Radiotherapy and Oncology* **2019**, *131*, 101–107. https://doi.org/10.1016/j.radonc.2018.10.037.
13. Li, L.; Zhao, X.; Lu, W.; Tan, S. Deep learning for variational multimodality tumor segmentation in PET/CT. *Neurocomputing* **2020**, *392*, 277–295. https://doi.org/10.1016/j.neucom.2018.10.099.
14. Zhao, X.; Li, L.; Lu, W.; Tan, S. Tumor co-segmentation in PET/CT using multi-modality fully convolutional neural network. *Physics in Medicine and Biology* **2019**, *64*. https://doi.org/10.1088/1361-6560/aaf44b.
15. Jin, D.; Guo, D.; Ho, T.Y.; Harrison, A.P.; Xiao, J.; kan Tseng, C.; Lu, L. DeepTarget: Gross tumor and clinical target volume segmentation in esophageal cancer radiotherapy. *Medical Image Analysis* **2021**, *68*. https://doi.org/10.1016/j.media.2020.101909.
16. Harrison, A.P.; Xu, Z.; George, K.; Lu, L.; Summers, R.M.; Mollura, D.J. Progressive and Multi-Path Holistically Nested Neural Networks for Pathological Lung Segmentation from CT Images. *International Conference on Medical Image Computing and Computer-Assisted Intervention* **2017**. https://doi.org/10.1007/978-3-319-66179-7_71.
17. Wang, H.; Hu, J.; Song, Y.; Zhang, L.; Bai, S.; Yi, Z. Multi-view fusion segmentation for brain glioma on CT images. *Applied Intelligence* **2022**, *52*, 7890–7904. https://doi.org/10.1007/s10489-021-02784-7.
18. Ma, L.; Chi, W.; Morgan, H.E.; Lin, M.H.; Chen, M.; Sher, D.; Moon, D.; Vo, D.T.; Avkshtol, V.; Lu, W.; et al. Registration-Guided Deep Learning Image Segmentation for Cone Beam CT-based Online Adaptive Radiotherapy. *Medical Physics* **(2022)**. https://doi.org/10.1002/mp.15677.
19. Fu, Y.; Lei, Y.; Wang, T.; Tian, S.; Patel, P.; Jani, A.B.; Curran, W.J.; Liu, T.; Yang, X. Pelvic multi-organ segmentation on cone-beam CT for prostate adaptive radiotherapy. *Medical Physics* **2020**, *47*, 3415–3422. https://doi.org/10.1002/mp.14196.
20. Lei, Y.; Wang, T.; Tian, S.; Dong, X.; Jani, A.B.; Schuster, D.; Curran, W.J.; Patel, P.; Liu, T.; Yang, X. Male pelvic multi-organ segmentation aided by CBCT-based synthetic MRI. *Physics in Medicine and Biology* **2020**, *65*. https://doi.org/10.1088/1361-6560/ab63bb.

21. Jia, X.; Wang, S.; Liang, X.; Balagopal, A.; Nguyen, D.; Yang, M.; Wang, Z.; Ji, J.X.; Qian, X.; Jiang, S. Cone-Beam Computed Tomography (CBCT) segmentation by adversarial learning domain adaptation. *Medical Image Computing and Computer Assisted Intervention* **2017**.
22. Brion, E.; Léger, J.; Barragán-Montero, A.M.; Meert, N.; Lee, J.A.; Macq, B. Domain adversarial networks and intensity-based data augmentation for male pelvic organ segmentation in cone beam CT. *Computers in Biology and Medicine* **2021**, *131*. https://doi.org/10.1016/j.compbiomed.2021.104269.
23. Zhao, S.; Lau, T.; Luo, J.; Chang, E.I.C.; Xu, Y. Unsupervised 3D End-to-End Medical Image Registration with Volume Tweening Network. *IEEE Journal of Biomedical and Health Informatics* **2019**. https://doi.org/10.1109/JBHI.2019.2951024.
24. Taha, A.A.; Hanbury, A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging* **2015**, *15*. https://doi.org/10.1186/s12880-015-0068-x.
25. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; Facebook, Z.D.; Research, A.I.; Lin, Z.; Desmaison, A.; Antiga, L.; et al. Automatic differentiation in PyTorch. *Neural Information Processing Systems Workshop* **2017**.
26. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *International Conference on Learning Representations* **2014**. https://doi.org/10.48550/arXiv.1412.6980.