3D Segmentation of Hard and Soft Tissue to Simulate X-ray Image Formation Using Deep Learning: Background Reading Report

Sean Darcy, Zhiyuan Ding, Qiyuan Wu

I. Introduction

In this report, we review background reading conducted for the project '3D Segmentation of Hard and Soft Tissue to Simulate X-ray Image Formation Using Deep Learning.' We begin by presenting DeepDRR [3]- the framework for X-ray image simulation that this project aims to improve via integration of a refined 3D segmentation module. Next, we review the literature surrounding 3D CT segmentation of multiple organ and tissue types, which was critically evaluated in the earlier phases of the project to design the refined 3D segmentation module. Finally, we present an article pertaining to a deep learning approach using X-ray images to solve a downstream clinical task. Comparing the performance of models trained on real vs. simulated data to achieve these 'downstream tasks' reviewed in the literature will serve to evaluate the quality of DRRs generated using our novel pipeline. In this report, details of the selected publications are summarized, rationale for selection is provided, and pros/cons in relation to our project are discussed.

II. Paper 1: DeepDRR–A Catalyst for Machine Learning in Fluoroscopy-guided Procedures

Authors: Mathias Unberath, Jan-Nico Zaech, Sing Chun Lee, Bastian Bier, Javad Fotouhi, Mehran Armand and Nassir Navab.

Reason for selection: DeepDRR is the CT-based X-ray image simulation pipeline that our project aims to improve. A comprehensive understanding of DeepDRR and its components is critical to our project's success.

Relevance to our work: The core of our project involves updating the current DeepDRR framework. Since the current version of DeepDRR only segments the input CT scan into bone, soft tissue and air, our project will focus on the materials segmentation component of the framework. Specifically, we aim to segment various additional categories of organs for a X-ray image and tissue types for more realistic X-ray image simulation.

Technical summary: Digitally reconstructed radiographs (DRRs) are simulated X-ray images. DRRs represent a promising strategy to improve the performance of artificial intelligence models aimed to augment or automate surgeon capabilities via generation of large simulated training datasets. A DRR is generated by 'imaging' a computed tomography (CT) volume *in silico*. Annotation and augmentation can be performed on the CT volume as opposed to individual images- reducing workload and promoting valid image characteristics. As suggested in [3], conventional machine learning based DRRs fail to generalize toward clinical application due to inaccurate model parameters under specific contexts. DeepDRR[3] was proposed for universal DRR generation with variable settings. In this way, users can obtain demanded DRRs by simply tuning parameters based on the necessary requirements.

DeepDRR is composed of four modules, as shown in Fig II, which include: (1) A V-Net based material decomposition module, where masks for various materials are generated with a combination of threshold strategy and network segmentation. (2) A material-and-spectrum-aware ray-tracing forward projector, where attenuation densities for different materials are used for X-ray spectral simulation at different positions. (3) A learning-based convnet- which consists of both Rayleigh scatter estimate and smoothness components- is used for scatter estimation using projection domain image results from part 2. (4) A noise generation module is constructed to simulate realistic noise injection. The amount of photons is estimated based on the spectrum image value and quantum noise is generated with the Poisson generating function.

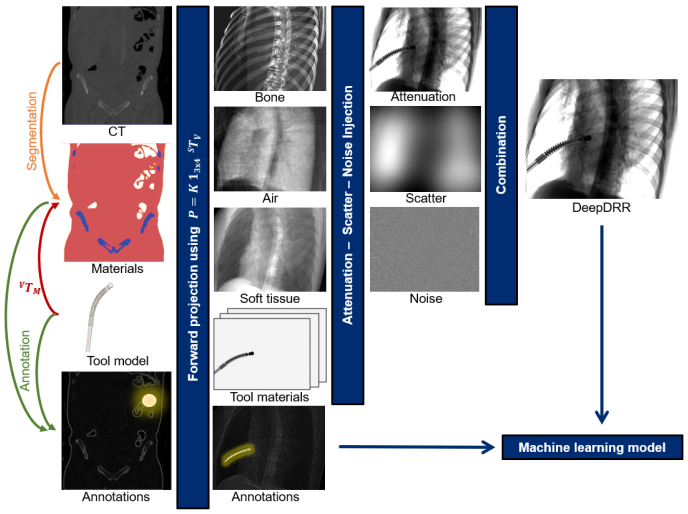

Fig. 1. The DeepDRR framework- which includes V-net based material decomposition, forward projector, neural network based scatter estimation, and noise generation components [3].

Pros and Cons, Takeaways:

Pros:

1. DeepDRR provides a general framework for CT based X-ray image simulation with useful toolboxes for simulation with tunable parameters. The modularized structure of the framework provides an excellent opportunity for module-specific updates, which our project aims to achieve.

Cons:

1. Previous work with DeepDRR constructed the general pipeline for CT-based X-ray image simulation. In the current DeepDRR release- bone, soft tissue, and air are each treated as different materials for spectrum attenuation- resulting in distinct appearances for each material in the simulated result. Currently, the DeepDRR materials segmentation stage classifies all non-bone tissue types as 'soft tissue,' with a single shared X-ray attenuation coefficient. A more realistic rendering would assign a specific X-ray attenuation coefficient to each tissue/organ type. Thus, there is a clear opportunity to render more realistic DRRs by augmenting the current CT segmentation stage of DeepDRR to include segmentations for a variety of organ and tissue types beyond just air, bone, and soft tissue- to which we will assign accurate, tissue-specific X-ray attenuation coefficients.

Takeaways:

1. We use the V-net part for air, bone and soft tissue segmentation to get bone and air mask for X-ray simulation.

2. We use the general X-ray image simulation process with potential updating in each toolbox.

III. Paper 2: nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentation [1]

Authors: Fabian Isensee, Jens Petersen, Andre Klein, David Zimmerer, Paul F. Jaeger, Simon Kohl, Jakob Wasserthal, Gregor Koehler, Tobias Norajitra, Sebastian Wirkert, and Klaus H. Maier-Hein.

Reason for selection: This paper presents detailed technical approach for nnUNet, along with executable repositories and pretrained models provided. This method suits well with the problem we wish to solve in DeepDRR system and with good performance as reported in the paper.

Relevance to our work: 3D segmentation from CT images is a critical component of the entire DeepDRR framework. With multiple regions corresponding to different tissue types segmented, unique parameters regarding tissue-X-ray interaction can be assigned in the projection simulation process, which would enable DeepDRR system to render simulated X-ray images with high fidelity. This paper presents a generalizable framework for 3D multi-organ segmentation, with accurate performance on multiple datasets and variety of tasks.

Technical summary: This paper mainly focuses on using varieties of the U-net- which is a specific class of artificial neural network architecture- to segment medical images. Several combinations of network architecture are described in this work- including 2D U-net, 3D U-net, U-net cascade and ensembled inference from multiple models. For different input datasets, each architecture is compared in terms of evaluation matrices (e.g. Dice score). Models with the best performances are selected, either alone or ensembled, for inference (Fig.III). By doing this, the architecture can actually be adjusted according to the dataset, and the performance would be improved compared to a model with fixed network architecture. Also, adaptive preprocessing (including cropping, resampling and normalization) and postprocessing are carried out in the repository provided in this paper. Finally, performance of this method is presented, with comparison over different organs, as is shown in Fig.III.

		2D U-Net	3D U-Net	3D U-Net lowres
BrainTumour	median patient shape	169x138	138x169x138	-
	input patch size	192x160	128x128x128	-
	batch size	89	2	-
	num pool per axis	5, 5	5, 5, 5	-
Heart	median patient shape	320x232	115x320x232	58x160x116
	input patch size	320x256	80x192x128	64x160x128
	batch size	33	2	2
	num pool per axis	6, 6	4, 5, 5	4, 5, 5
Liver	median patient shape	512x512	482x512x512	121x128x128
	input patch size	512x512	128x128x128	128x128x128
	batch size	10	2	2
	num pool per axis	6, 6	5, 5, 5	5, 5, 5
Hippocampus	median patient shape	50x35	36x50x35	-
	input patch size	56x40	40x56x40	-
	batch size	366	9	-
	num pool per axis	3, 3	3, 3, 3	-
Prostate	median patient shape	320x319	20x320x319	-
	input patch size	320x320	20x192x192	-
	batch size	26	4	-
	num pool per axis	6, 6	2, 5, 5	-
Lung	median patient shape	512x512	252x512x512	126x256x256
	input patch size	512x512	112x128x128	112x128x128
	batch size	10	2	2
	num pool per axis	6, 6	4, 5, 5	4, 5, 5
Pancreas	median patient shape	512x512	96x512x512	96x256x256
	input patch size	512x512	96x160x128	96x160x128
	batch size	10	2	2
	num pool per axis	6, 6	4, 5, 5	4, 5, 5

Fig. 2. Network topologies as automatically generated for seven tasks. 3D U-Net lowres refers to the first stage of the U-Net Cascade [1].

	BrainTumour			Heart	Liver		Hippoc.		Prostate		Lung	Pancreas	
label	1	2	3	1	1	2	1	2	1	2	1	1	2
2D U-Net	78.60	58.65	77.42	91.36	94.37	53.94	88.52	86.70	61.98	84.31	52.68	74.70	35.41
3D U-Net	**80.71**	**62.22**	**79.07**	92.45	94.11	61.74	**89.87**	**88.20**	60.77	83.73	55.87	77.69	42.69
3D U-Net stage1 only (U-Net Cascade)	-	-	-	90.63	94.69	47.01	-	-	-	-	65.33	79.45	49.65
3D U-Net (U-Net Cascade)	-	-	-	92.40	95.38	58.49	-	-	-	-	**66.85**	**79.30**	**52.12**
ensemble 2D U-Net+ 3D U-Net	80.79	61.72	79.16	**92.70**	94.30	60.24	89.78	88.09	**63.78**	**85.31**	55.96	78.26	40.46
ensemble 2D U-Net+ 3D U-Net (U-Net Cascade)	-	-	-	92.64	95.31	60.09	-	-	-	-	61.18	78.79	45.46
ensemble 3D U-Net+ 3D U-Net (U-Net Cascade)	-	-	-	92.63	**95.43**	**61.82**	-	-	-	-	65.16	79.70	49.14
test set	67.71	**47.73**	**68.16**	**92.77**	**95.24**	**73.71**	**90.37**	**88.95**	**75.81**	**89.59**	**69.20**	**79.53**	**52.27**

Fig. 3. Mean dice scores for the proposed models in all tasks. All experiments were run as five-fold cross-validation. The models used for generating test set submission are highlighted in bold. The dice scores of the test sets are shown at the bottom of the table [1].

Pros and Cons, Takeaways:

Pros:

1. This paper presents an adaptable framework for medical image segmentation, which is generalizable for different organ

and tissue types.

2. This paper provides automated preprocessing and post-processing, which can be directly used on the dataset we are interested in.

3. This paper illustrates the detailed technical approach, which enable us to better understand the mechanism of our segmentation.

Cons:

1. The performance on lung segmentation isn't perfect, which still does not solve the current bottleneck of multi-organ segmentation.

2. Some of the pretrained models are trained on MR images rather than CT images, making it hard to directly infer from the input of CT images that we currently have.

Takeaways:

1. We could used different models for the segmentation of different organs since each model has its best performance on different tasks.

2. Heart and liver segmentation performed well by this approach, so we could integrate heart and liver segmentation onto DeepDRR to study if this segmentation process improves DRR simulation performance.

3. We combined the results from this paper and put together the values from other papers, and conclude in table I for a overall comparison of bench-marking approaches for multi-organ 3D segmentation.

TABLE I
REVIEW ON MULTI-ORGAN 3D SEGMENTATION

	nnU-Net	R2U3D	3D U-net	Hierarchical 3D-FCN	2D-FCN	3D-Deep-CNN
Lung	0.6920	0.9920	-	-	93.9(R) 93.5(L)	95.1(R) 94.4(L)
Heart	0.9277	-	0.8432	-	86.0	89.0
Liver	0.9524	-	-	0.954	90.8	91.1
Spleen	-	-	-	0.928	83.5	91.3
Heart	0.7581	-	-	-	47.0	74.2

IV. PAPER 3: LUNG NODULE DETECTION IN CHEST X-RAYS USING SYNTHETIC GROUND-TRUTH DATA COMPARING CNN-BASED DIAGNOSIS TO HUMAN PERFORMANCE [1]

Authors: Manuel Schultheiss, Philipp Schmette, Jannis Bodden, Juliane Aichele, Christina Müller-Leisse, Felix G Gassert, Florian T Gassert, Joshua F Gawlitza, Felix C Hofmann, and Daniel Sasse.

Reason for selection: Evaluating the quality of simulated data is challenging. Here, we define improved 'quality,' as increased clinical applicability of simulated data due to increased realism in the simulation process. To evaluate 'clinical applicability,' we have reviewed several downstream tasks that make use of X-ray image data relevant to the organs and tissues for which our novel 3D segmentation module provides segmentations. Our downstream evaluation task will deploy a nearly identical approach as in [2] described below, with the main difference being the use of two versions of DeepDRR (current and updated) to generate the simulated data.

Relevance to our work: To evaluate whether the addition of a refined 3D CT segmentation module to the DeepDRR pipeline improves DRR quality, we will compare performance of downstream task models tested on real X-ray images, but trained on either real X-ray images, current DeepDRRs, or novel DeepDRRs. This strategy aims to measure the clinical applicability of our simulated data. Improved DRR quality would manifest as improved testing performance of the downstream task model trained on novel DeepDRRs as compared to the same model architecture trained with current DeepDRRs. Approaching or even surpassing (with larger simulated datasets) downstream task test performance of a model trained with real X-ray images is the goal.

Technical summary: For the lung, we found nodule detection to be a popular task in the literature. [2] trained a U-Net and supplementary lesion scoring network using synthetic data (DRRs) and evaluated performance on real X-ray images against trained radiologists.

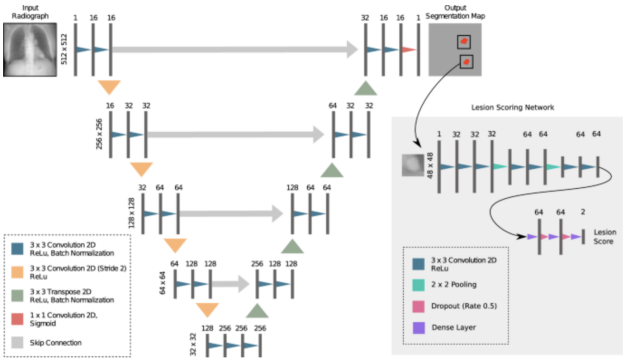

Fig. 4. U-Net and supporting lesion scoring network for lung nodule detection downstream evaluation task [2].

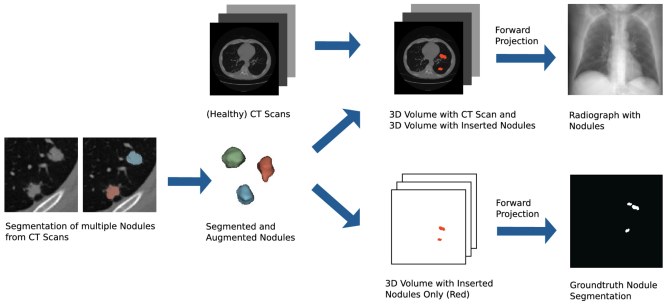

Fig. 5. Data generation pipeline for lung nodule detection downstream task [2].

Here, the synthetic data was generated by first obtaining nodule segmentations from diseased lungs, augmenting/perturbing these nodule segmentations, then randomly inserting them into healthy lung CTs. The digitally reconstructed

p.a. thorax radiographs are then generated via forward projection of the CTs. In parallel, ground truth masks are generated via forward projection of just the nodule segmentations, as in Fig. 5.

Pros and Cons, Takeaways:

Pros:

1. This paper provides a generalizable process for evaluating the clinical applicability of simulated X-ray data. This nodule detection task may be applied to additional organs segmented in our updated DeepDRR framework, such as the liver.

2. In addition to evaluating performance of the downstream network described for both current and updated versions of DRR, we may compare against clinician performance given the same test set of real X-ray images, without having to acquire the clinician performance data ourselves.

Cons:

1. As discussed with our mentors, 'sim-to-real' effects which currently limit the clinical applicability of simulated data may 'overshadow' any improvements that our 3D segmentation module may provide. This would manifest as little-to-no improvement in the performance of the updated DeepDRR-trained U-net + lesion scoring network as compared to that of current DeepDRR-trained network.

2. An additional pre-processing step is required to insert nodules into healthy CT scans for ground truth generation. This requires more compute than simpler tasks such as landmark detection in bones which has been used to evaluate DeepDRR previously [3].

Takeaways:

1. We will start with a lung nodule segmentation task as inspired by [2] to evaluate the improved clinical applicability of our updated DeepDRRs.

2. We will generalize this nodule segmentation task to other organs segmented in our updated DeepDRR pipeline, such as the liver and kidneys.

3. In addition, we will continue to evaluate our simulated images via bone landmark detection task, as in [3].

3. We are well-aware of sim-to-real challenges in the lung nodule detection task that may occlude any improvement that we are able to make to the rendered DRRs. We have discussed options to mitigate these challenges. In any case, bone landmark detection has proven to be an effective downstream evaluation task.

References

[1] Fabian Isensee et al. "nnu-net: Self-adapting framework for u-net-based medical image segmentation". In: *arXiv preprint arXiv:1809.10486* (2018).

[2] Manuel Schultheiss et al. "Lung nodule detection in chest X-rays using synthetic ground-truth data comparing CNN-based diagnosis to human performance". In: *Scientific Reports* 11.1 (2021), pp. 1–10.

[3] Mathias Unberath et al. "DeepDRR–A Catalyst for Machine Learning in Fluoroscopy-guided Procedures". In: *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*. Springer, 2018.