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

Zhiyuan Ding, Qiyuan Wu, Sean Darcy

I. Background

X-ray images have been widely used to guide clinical decision making for over a century. More recently, artificial intelligence (A.I.) models trained on X-ray image data have demonstrated promising results on specific constrained tasks such as bone fracture detection, anatomical landmark detection, and detection of certain tumors and abnormal masses. Previous research has demonstrated the advantage of X-ray in tasks such as these over images acquired through other modalities [1], [2].

Fluoroscopy-guided interventions are minimally-invasive surgical procedures during which the surgeon is guided by a continuous stream of X-ray images displayed on a monitor. This class of intervention has become popular in orthopedic surgeries, liver biopsies, pacemaker implantations, catheter insertions, angiography, and a variety of other procedures. Fluoroscopy-guided intervention is a rapidly growing discipline and one that is prime for automation- particularly using deep learning approaches with models trained on intraoperative image data

Successful approaches to automation of fluoroscopy-guided interventions require substantial amounts of high quality training data, but this data is difficult to obtain. As pointed out in [3], X-ray images, especially in 2D fluoroscopy-guided procedures, are rarely stored in a meaningful way and learning targets are not widely established. At the same time, data that is stored is of high variability, since different tools may be present in different frames of varying viewpoints. This makes scalable annotation of high-quality ground truth fluoroscopy data practically impossible.

One potential solution to this problem is simulating X-ray images in silico from diagnostic 3D Computed Tomography (CT). Annotation and augmentation can be performed on the CT volume as opposed to individual images- reducing workload and promoting valid image characteristics. Several works have previously utilized CT as input for X-ray image generation, such as [3]–[5]. Until recently, these simulated X-ray images were not very realistic and failed to translate to the clinic.

DeepDRR [3] is an X-ray image simulation pipeline developed here at Hopkins and provides state of the art tools to generate realistic radiographs at training set scale. It renders the most realistic DRRs to date as demonstrated by a pelvis landmark detection task, where DeepDRR radiographs substantially outperformed other DRRs. Despite being more realistic, current DeepDRR images are still not indistinguishable from true radiographs.

II. Problem statement

Our work primarily consists of updates to the previous DeepDRR pipeline [3] aimed at improving the quality of simulated images. The first stage in the DeepDRR pipeline consists of a materials segmentation through which material coefficients are later assigned. Based on material categories assigned to CT voxels, ray-tracing for image generation and scattering are processed in later DeepDRR stages. Limitations to the previous DeepDRR pipeline and our corresponding deliverables are as follows:

- First, the quality of segmented materials masks should be enhanced. After analysis of output from previous DeepDRR material masks, which are segmented via combination of V-net and thresholding methods, we find that key anatomical details are not rendered properly in the generated X-ray simulation result. Thus, in order to improve the simulation result, more accurate segmentation masks are required.
- Second, the available categories of organs and tissue types should be expanded. One major limitation to the previous DeepDRR pipeline was the limited number of materials assigned. Only bone, soft tissue, and air materials coefficients were assigned based on thresholding methods. These simulated radiographs suffered from sim-to-real challenges limiting their applicability in downstream clinical A.I. training schemes. In reality, there are more than just 3 classes of material in the human body, and each has its own unique properties with respect to interaction with X-rays. To render more realistic DRRs, a broader variety of materials and their properties should be accurately represented in our in silico projections.
- Finally, the improved quality of X-ray image simulation should be evaluated quantitatively. Improved quality can be defined in this context as increased clinical applicability of A.I. trained with the simulated data, measured by testing performance in the real domain. DeepDRR quality was previously evaluated on a bone landmark detection task. Here, we aim to illustrate the utility of expanded and increasingly accurate tissue/organ representations in simulating datasets for applications across a variety of clinically relevant tasks. The real-image testing

performance of downstream A.I. trained on sets of either previous or novel DeepDRR simulated data is compared. We present a lung nodule detection task to achieve this evaluation, which may be generalized to nodule detection tasks for a variety of organs.

In this report, the approaches for each DeepDRR update and deliverables are introduced. In section 3, we describe our novel 3D segmentation module consisting of a set of high performing pretrained 3D CT segmentation models that were reviewed, evaluated and deployed to achieve more accurate and detailed mask generation. Additionally, we describe the merging process for masks output from our chosen set of pretrained models. We proceed to describe the design of a clinically-relevant downstream evaluation task to quantitatively measure improved X-ray image simulation quality. In section 4, several key results are presented- including 3D segmentation performance for abdominal CT organ masks rendered by our novel segmentation module, integration of our novel segmentation module to the DeepDRR pipeline, example DRR images for visual comparison, and example images from the dataset generated for downstream task evaluation. In section 5, we discuss project conclusions and significance.

III. Approaches

The general project framework is shown in Figure 1. First, input CT images are segmented into various organs by different models. Then the obtained masks for a specific case are merged to form the final multi-organ mask. The obtained mask is then assigned materials coefficients to be referenced for DeepDRR X-ray image simulation. Finally, a downstream experiment is deployed to evaluate the simulation result.

Our work for updating previous version DeepDRR can be categorized into three aspects: First, considering both the performance and efficiency of total pipeline, masks for various organs are extracted based on models pre-trained on existing datasets. Then, the masks are collected and combined to generate the merging masks for DeepDRR input. Finally, with DeepDRR generation, a downstream task is proposed for simulation evaluation.

A. Pretrained models used for segmentation

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. In this work, two models trained for various organ segmentation were utilized for mask generation.

1) nnUNet: Residual Networks are widely used in many tasks of medical image segmentation, and various architectures are developed including U-net, V-net, etc.. [6] Here a ResNet framework called 'nnU-Net' [7] was applied in our project, for the segmentation of abdominal CT. NnU-Net is a multi-architecture deep learning framework for image segmentation, which automatically adjusts the architecture and other settings according to the input training data, and the pretrained models attached are proved to have promising performance on some medical image segmentation tasks. Multiple pretrained models from nnUNet repository were applied to our pipeline respectively, and users can choose the one they prefer in the pipeline we developed. In the following part of this report, the results and discussions are all based on the pretrained model of 'Task017_AbdominalOrganSegmentation'.

2) CT-ORG Net: CT-ORG net is a U-net based model trained on CT-ORG dataset [8]. In this dataset the CT volumes are annotated for 6 categories. Here in our task, the trained model is used for organ mask generation, especially for bone segmentation.

B. Label Fusion and Multi-organ Mask Generation

With the obtained mask output from different models, in this stage a merging process is deployed to generate the final mask as DeepDRR input. This merging process consists of two steps. First, masks of different organs are collected into one space. If different models contains masks of the same type of organs, we select the one with organ mask in its neighboring area to avoid overlapping of different organs. If the segmented organ in a position with no neighboring organ, then the one with better performance is selected. Second, since voxel labels in DeepDRR are required to work as 'density' for attenuation simulation, a threshold segmentation process is used to classify remaining voxels into air and general soft tissue categories. This process is done by a comparison with estimated distribution of air and general soft tissue voxel values in CT data.

C. Downstream Task Method

The ultimate goal of integrating a novel 3D segmentation pipeline with DeepDRR is to improve the quality of simulated X-ray (DRR) datasets. Quality in this context is defined as increased (real) clinical applicability of A.I. models trained on simulated data. To assess improved DRR quality, we compare the testing performance on real X-ray images of downstream-task models trained on either DeepDRRs generated with our novel 3D segmentation module or DeepDRRs as they are rendered today (Fig. 3). These models must learn from input X-ray images in order to achieve some downstream clinical task. We would expect higher quality, more clinically applicable simulated training data to result in higher performing downstream A.I. performance in the real domain (Fig. **??**).

We selected nodule detection as a clinically-relevant downstream task that would enable us to quantitatively evaluate the improved quality of our DRRs. We chose nodule detection due to its generalizability across the organ and tissue types for which our novel 3D segmentation module provides masks.

IV. Results

In this section, experiment details and corresponding results are shown. The project deliverables consist of three parts. First, we illustrate the performance of masks output from different models. Model performance was evaluated on the Pediatric

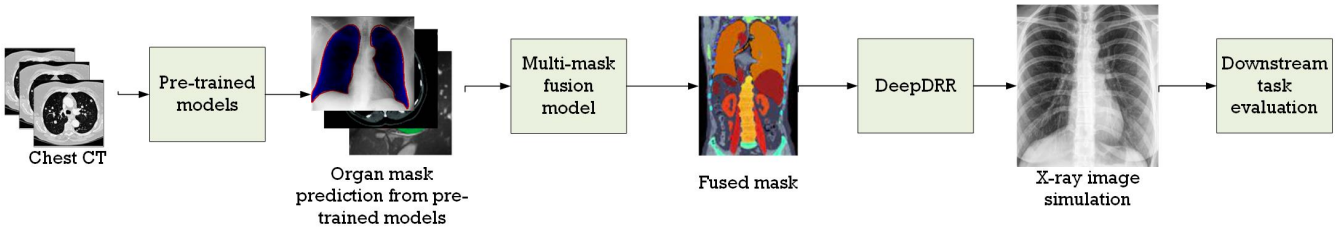


Fig. 1. General project framework

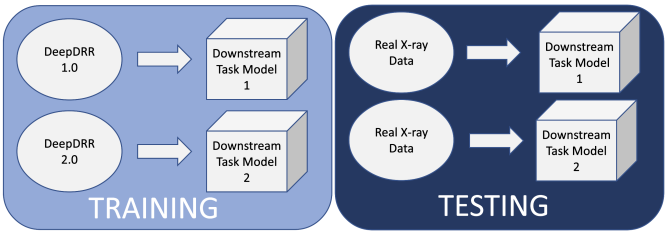


Fig. 2. 2 Models are trained with separate sets of simulated data (old and new), but tested with the same set of real data to quantitatively evaluate simulated data quality for learning.

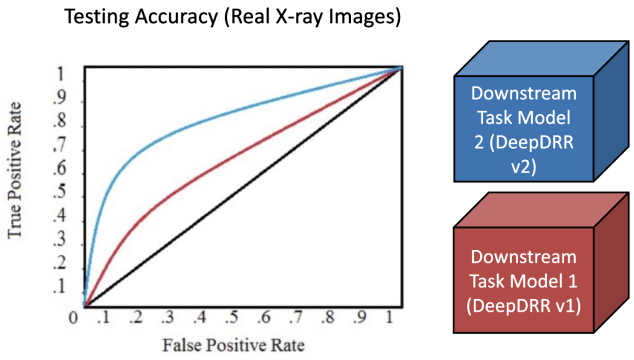


Fig. 3. Downstream, simulated-data-trained model real image accuracy will be compared quantitatively. We expect the model trained with our novel DeepDRR-rendered datasets will have higher performance.

[9] dataset. Next, X-ray image simulation results generated from 98 high resolution full-body adult CT images from The New Mexico Decedent Image Database (NMDID) [10] are displayed and analyzed. Finally the details of downstream task are provided.

A. Segmentation for abdominal CT

In this project, nnUNet and CT-ORG Net are used to generate masks for different organs. These models contain parameters trained from their respective original datasets. The segmentation models are evaluated on the Pediatric dataset [9], which contains 350 cases of children abdominal CT images. In this experiment, 150 random selected cases are used for test. The experiment result is shown in Table I. Dice score, the ratio of overlapping of segmentation mask and ground truth, is used as measurement for this test. Pediatric is the only dataset we find with a substantial amount of multi-organ annotated cases. So, although the dataset only contains CT images for children, we use it for this preliminary mask evaluation to check the model can at least locate the organs in their corresponding anatomical locations.

Model	Organ	Dice-Score
nnUNet	spleen	0.6358
	kidney right	0.5954
	kidney left	0.5814
	gallbladder	0.3782
	esophagus	0.3265
	liver	0.6948
	stomach	0.4411
	pancreas	0.3069
CT-ORG Net	lung	0.6429
	liver	0.7142
	kidney	0.5675
	bone	0.7326
	bladder	0.4083

TABLE I
SEGMENTATION RESULT FOR NNUNET AND CT-ORG NET ON PEDIATRIC DATASET.

Based on the performance shown in Table I, spleen, gallbladder, esophagus, and stomach are selected from nnUNet and all of the five categories in CT-ORG Net are selected. The selected organ masks are used as the basis for label fusion in the next step.

B. Label Fusion and Multi-organ Mask generation

Considering both the performance and its actual spatial distribution, masks from nnUNet and CT-ORG Net are collected in the following categories: In nnUNet, spleen, gallbladder, esophagus, and stomach are selected and in CT-ORG Net lung, liver, kidney, bone and bladder are selected.

After that ,the threshold segmentation is deployed to classify remaining voxels into air or general soft tissue. We use coefficient for this threshold method the same as in [3], which assume the mean of CT image value is -630.1 and standard deviation is 479.7. Classification threshold of air and general soft tissue is -500.

The mask result is shown in Figure 4. Details about bone and and organs are delineated in the mask. Further evaluation

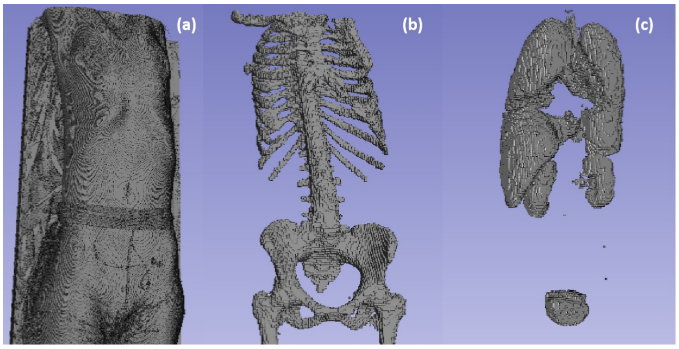


Fig. 4. Generated mask for multiple organs or tissues. (a) An example of air and general soft tissue mask. (b) An example of bone segmentation mask. (c) An example of segmented organs' mask.

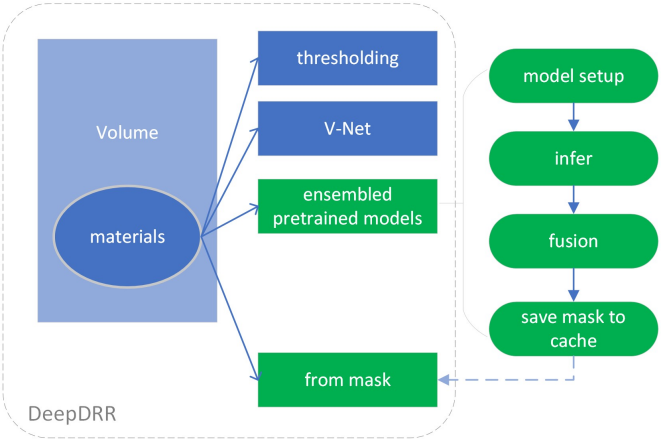


Fig. 5. Development in DeepDRR: Light Blue: existing class; Blue: existing modules for assigning segmented materials; Green: Newly developed modules. The 'from mask' module reads the corresponding segmentation of the input CT image and assigns segmentation to 'materials'. The 'ensembled pretrained models' module sets and calls the pretrained model from nnU-Net and CT-ORG, fuses the output masks altogether with thresholding as supplementation for those unlabeled pixels, and saves the fused mask into a temporary cache, finally uses the 'from mask' module to read the mask and assign to 'materials'.

of the performance of segmentation result can be seen from the generated X-ray simulation image.

C. Integration of segmentation pipeline into DeepDRR

One of the goals of this project was to develop the DeepDRR package to contain all of the described segmentation pipelines within a single parameterized module, so that user can perform DRR simulation with refined segmentation automatically, without needing to prepare the refined segmentation masks themselves. Figure 5 shows the improvements made on DeepDRR package, which includes utilities for mask preparation, label fusion and linking to the external packages. A label fusion algorithm is developed. This fusion process consists of a multi-organ mask choice and a threshold classification for unlabeled pixels. Several internal utility modules are developed to prepare, set and call the external packages and pretrained models, and intensive path management is performed for linking to external pretrained models. To reduce the complexity and the size of DeepDRR package, the pretrained models are stored and called from external source. [11]

Simulated X-Ray images are generated from the new pipeline using fused mask as the segmentation method, along with the comparison to the old ones generated using thresholding as the segmentation method, as shown in Fig.6 and Fig.7, for pelvis and chest respectively. For pelvis images, the results with fused mask have an enhanced intensity contrast between bone and other tissues, but lose some resolution compared to the old one with thresholding. This may not necessarily be a drawback, the new X-Ray image could be more realistic since the real X-Ray scans may not have such a high resolution as those simulated images with thresholding. The enhancement of contrast intensity leads to a clearer edge especially at femur head as well as pubic arch. Similar chances happen in the comparison of chest images. The intensity of vertebrae is enhanced while the resolution reduces.

Currently, only 4 kinds of mass attenuation coefficient are assigned (bone, air, lung, and soft tissue (including kidney, liver, etc.)) on our 11-channel segmentation, but those comparison in Fig.6 and Fig.7 already show some visual changes regarding the bone part, possibly reflecting the improvement of bone segmentation since the fixed threshold values in the old method do not perform well as the input image varies. With the current outcome, more improvements can be expected in the very near future when we treat different organs with their unique mass attenuation coefficients.

D. DeepDRR performance evaluation by downstream tasks

To assess improved DRR quality, we compare the testing performance on real X-ray images of downstream-task models trained on either DeepDRRs generated with our novel 3D segmentation module or DeepDRRs as they are rendered today. We began with lung nodule detection, with methods inspired by [12]. First, we generated a library of 5 nodules segmented from the LUNA-16 dataset [13]. Then, for each CT volume (87 in total) in our NMDID [10] subset, we inserted between 0 and 3 nodules randomly selected from the library. Each of the nodules was randomly rotated, randomly scaled from .2 up to 1.1, and placed at a random point within the organ of interest- in this case, the lung. Some examples are displayed in Fig. 8 and Fig. 9. We have generated 15 DRRs per each of 87 available CT scans. For each CT, we randomly selected, perturbed, and inserted between 0 and 3 nodules a total of three times, and each of these was 'imaged' in silico from 3 distinct views. Out-of-frame images were discarded. Both old and novel versions of the DeepDRR simulation were run for each case, insertion instance, and projector view. Thus, we were able to render two datasets of 774 distinct DRRs with identical groundtruth.

We chose the segmentation-state-of-the-art U-Net [14] architecture as our downstream lung-nodule detection model architecture, as in [12]. Data augmentation included rotation and flip operations.The U-Net was trained for 50 epochs and a batch size of 1. Adam optimizer parameters were set to β_1

Fig. 6. Simulated X-Ray images for pelvis. Left: Fused mask as segmentation method; Right: Thresholding as segmentation method. Upper Row: AP view; Middle Row: Lateral view; Bottom Row: Obturator oblique view. The red circles show the region where contrast of bone is significantly enhanced.

= 0.9 and β_2 = 0.999. Learning rate was set to 1.5. The loss function used to train was Dice loss.

We separately trained two instances of the same downstream model- one with the thresholding-based DeepDRR rendered dataset and one with the novel, segmentation-based version. After training, we evaluated the performance of each via comparison of the DICE accuracy score obtained from a dataset of real input X-ray images with 0 to 3 nodules. Testing these models on only real X-ray images and comparing performance allows us to validate that our improved DRRs effectively train models meant to operate on real, non-simulated data generated in the clinic. We expect to observe improved testing accuracy for models trained on DeepDRRs generated with our novel 3D segmentation module compared to models trained on DeepDRRs rendered via thresholding.

For training, we observed no decreasing loss (see Fig. 10) for the U-Net trained on the old DeepDRR data, which

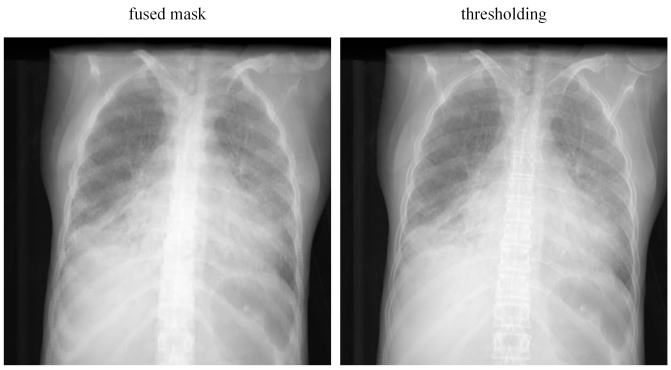


Fig. 7. Simulated X-Ray images for chest in AP view. Left: Fused mask as segmentation method; Right: Thresholding as segmentation method.

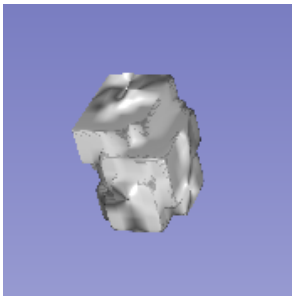


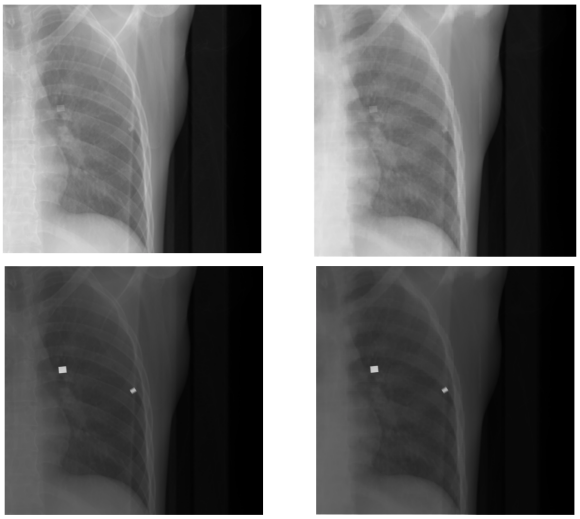
Fig. 8. Example of nodule segmentation to be inserted in CT volume prior to imaging in silico. Each nodule was randomly rotated, scaled, and inserted randomly within the organ of interest.

fluctuated at a loss over .99. For the U-Net trained on new DeepDRR data with our segmentation module, we observed convergence at just over .75 loss after 10 epochs. This could indicate that the segmentation model was able to learn better via a stabler optimization landscape created by our new data. However, this could also indicate overfitting, where the model essentially memorizes the training examples, as opposed to learning meaningful features for nodule segmentation.

For testing, we manually annotated 5 real X-ray images from the JSRT dataset [15]. The DICE score for the old DeepDRR-trained model was 0.0148, while that of the new DeepDRR-trained model was 0.0315. These results are nowhere close to what is required for deployment in a clinical setting. Still, a higher DICE score- which measures the overlap between segmentation masks and model output- for the model trained on data from our updated compared to the former version of DeepDRR is a promising result in that models trained on the updated simulated data perform better on real data in the clinic.

V. Significance and Discussion

The goal of this project was to update DeepDRR, a state-of-the-art X-ray image simulation pipeline, to render simulated images with increased clinical applicability. First, we improve DeepDRR segmentation performance by introducing a novel multi-organ segmentation module designed with carefully selected, high performing pre-trained models, in addition to



A) Thresholding-Based DeepDRR B) 3D Segmentation-Based DeepDRR

Fig. 9. DeepDRRs rendered for lung nodule detection task with overlaid masks

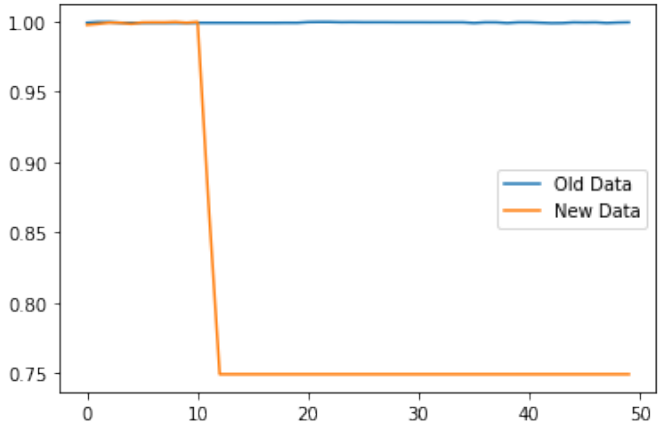


Fig. 10. Loss curves over 50 epochs for lung nodule segmentation models trained on simulated data rendered by either former or updated versions of DeepDRR.

expanding the organ categories to which materials coefficients are assigned. Second, a downstream task is proposed to evaluate the improved quality of the simulation result.

The generated result is improved and more key image details can be found in the updated version of DeepDRR, as we've shown in Figure 6. Additionally, for a specific organ target, we provide an illustration of how a specific organ segmentation model can be used to generate a more accurate segmentation mask for this organ and thus enhance the visually-inspected quality of corresponding simulation image.

For the second contribution, we provide a standard example for how to quantitatively evaluate the performance of X-ray simulation images based on downstream task A.I. performance evaluation. In this example, we developed a nodule insertion simulation framework, generalizable to any segmented organ in our updated DeepDRR pipeline. We found that a U-Net segmentation model trained on our updated segmentation-based simulated data outperformed its counterpart trained on the former, thresholding-based simulated data. Still, sim-to-real challenges remain, as the results were far from what is required in the clinic. Perhaps with a larger CT dataset, we would be able to render a larger X-ray image dataset which would be less prone to overfitting. Additionally, we might try expanding our nodule library to include more nodule shapes. Finally, lung nodule segmentation in X-ray is no easy task, even with real data, as nodules in X-ray are not easily visible as compared to CT. To further test whether our updated simulated data is of greater clinical applicability, we will use a similar approach described in the methods section for a variety of other tasks- including but not limited to nodule segmentation in other organs, organ segmentation generally, and anatomical landmark detection.

For future development, more pre-trained models can be added to our 3D segmentation module to further expand the available categories of organs and tissue types, as well as to improve the segmentation accuracy of existing categories. Additionally, for regions where the segmentation masks are not accurate enough, fuzzy boundaries may be used to combine probabilities from various categories. More downstream tasks, such as general organ segmentation in X-ray and anatomical landmark detection, should also be evaluated to confirm whether this simulated "quality" holds across a variety of clinical applications.

APPENDIX

For project code and documentation, please refer to https://github.com/wu-qiyuan/deepdrr. New Developments of this project are in *CIS-development* folder of branch *final-presentation*.

REFERENCES

[1] M. Haber, A. Drake, and J. Nightingale, "Is there an advantage to using computer aided detection for the early detection of pulmonary nodules within chest x-ray imaging?" *Radiography*, vol. 26, no. 3, pp. e170–e178, 2020.

[2] V. L. Parker, M. C. Winter, E. Whitby, W. A. Parker, J. E. Palmer, J. A. Tidy, A. A. Pacey, B. W. Hancock, and R. F. Harrison, "Computed tomography chest imaging offers no advantage over chest x-ray in the initial assessment of gestational trophoblastic neoplasia," *British Journal of Cancer*, vol. 124, no. 6, pp. 1066–1071, 2021.

[3] M. Unberath, J.-N. Zaech, S. C. Lee, B. Bier, J. Fotouhi, M. Armand, and N. Navab, "Deepdrr–a catalyst for machine learning in fluoroscopy-guided procedures," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2018, pp. 98–106.

[4] A. Maier, H. G. Hofmann, M. Berger, P. Fischer, C. Schwemmer, H. Wu, K. Müller, J. Hornegger, J.-H. Choi, C. Riess *et al.*, "Conrad—a software framework for cone-beam imaging in radiology," *Medical physics*, vol. 40, no. 11, p. 111914, 2013.

[5] L. Levine and M. Levine, "Drrgenerator: A three-dimensional slicer extension for the rapid and easy development of digitally reconstructed radiographs," *Journal of Clinical Imaging Science*, vol. 10, 2020.

[6] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, "A review of deep learning based methods for medical image multi-organ segmentation," *Physica Medica*, vol. 85, pp. 107–122, 2021.

[7] F. Isensee, J. Petersen, A. Klein, D. Zimmerer, P. F. Jaeger, S. Kohl, J. Wasserthal, G. Koehler, T. Norajitra, S. Wirkert, and K. H. Maier-Hein. [Online]. Available: https://arxiv.org/abs/1809.10486

[8] B. Rister, D. Yi, K. Shivakumar, T. Nobashi, and D. L. Rubin, "Ct-org, a new dataset for multiple organ segmentation in computed tomography," *Scientific Data*, vol. 7, no. 1, pp. 1–9, 2020.

[9] P. Jordan, P. M. Adamson, V. Bhattbhatt, S. Beriwal, S. Shen, O. Radermecker, S. Bose, L. S. Strain, M. Offe, D. Fraley *et al.*, "Pediatric chest-abdomen-pelvis and abdomen-pelvis ct images with expert organ contours," *Medical Physics*.

[10] S. D. Berry and H. J. Edgar, "Announcement: The new mexico decedent image database," *Annotation*, 2021.

[11] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnu-net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, no. 2, pp. 203–211, 2021.

[12] M. Schultheiss, P. Schmette, J. Bodden, J. Aichele, C. Müller-Leisse, F. G. Gassert, F. T. Gassert, J. F. Gawlitza, F. C. Hofmann, D. Sasse *et al.*, "Lung nodule detection in chest x-rays using synthetic ground-truth data comparing cnn-based diagnosis to human performance," *Scientific Reports*, vol. 11, no. 1, pp. 1–10, 2021.

[13] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. Van Den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge," *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[15] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-i. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," *American Journal of Roentgenology*, vol. 174, no. 1, pp. 71–74, 2000.

Management Summary

Sean Darcy, Zhiyuan Ding, Qiyuan Wu

May 2022

1 Work division

Qiyuan's Work:1. Exploration, Implementation, Modification of nnU-Net 2. Development on DeepDRR package: developed several utilities that automatically generate segmentation from nnUNet, read fused masks as material segmentation 3. X-Ray image generation: Adjusted, modified settings of DeepDRR to generate X-Ray image in a preferred view.

Zhiyuan's Work: 1. Implementation of MedicalNet for 3D segmentation with training(not applied in this version) and construction of merging network(not applied in this version). 2. Model evaluation of nnUNet and CT-ORG Net, implementation of CT-ORG Net. 3. Multi-organ mask generation for DeepDRR: batch-wise label fusion for nnUNet and CT-ORG Net.

Sean's Work: 1. Exploring CT-ORG Net and help implementing it for segmentation. 2. Design and implement downstream task for X-ray simulation image evaluation.

2 Planned and accomplished work (deliverable and timeline)

For code deliverable and documentation, please refer to https://github.com/wu-qiyuan/deepdrr. New Developments of this project are in *CIS-development* folder of branch *final-presentation*.

Deliverable (Expected date)	**Task**	**Status**
Minimal (March.8)	1. Scripts of a validated segmentation pipeline for bone, soft tissue, and air in full-body CT. 2. Comparison report of existing models on different type of tissues.	Completed Completed
Expected: (April.6)	1. Scripts and results for segmenting lung, liver in full-body CT. 2. Integrated scripts into the DeepDRR simulation.	Completed (modified goal, not include cardiac anymore but include other organs) Completed Further improvement is planned afterwards.
Maximum (May.3)	1. Scripts and results for segmenting multiple organs (including liver, spleen, kidney, etc.) in full-body CT. 2. Analysis report of the DRR performance using the improved segmentation pipeline	Completed Completed, with future goals planned

Table 1: Plan and accomplished Task

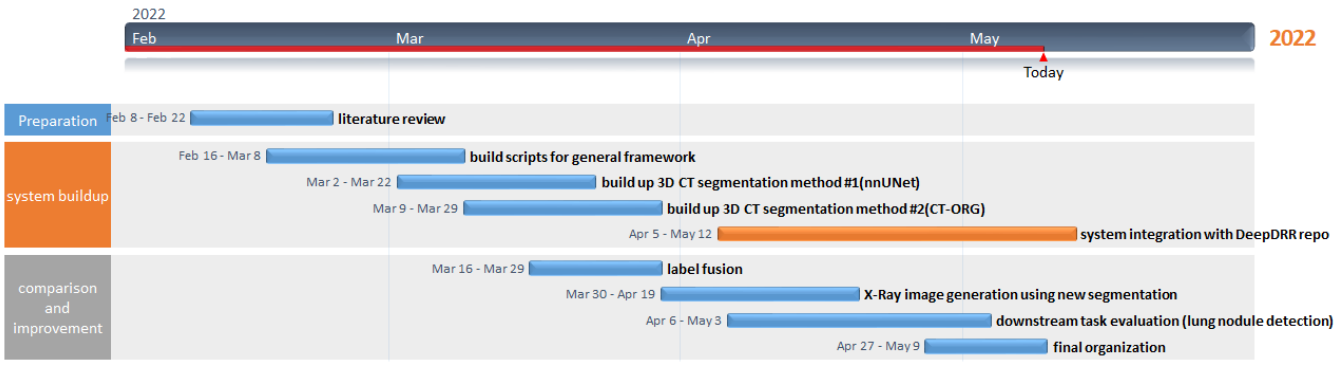


Figure 1: Project timeline. Yellow for ongoing task.

3 Future development

1. For task-specific X-ray image simulation, models trained on target specific dataset can be used to generate more accurate mask and some efforts to introduce these models can further improve the performance of DeepDRR.

2. Currently we assign a unique label for each voxel, and if the generalization ability of the segmentation model is not good enough, this would leads to some inaccurate results. Next step this mask can be replaced by a fuzzy mask for each voxel and considering the geometry distribution of each organ to make a better estimation of values on the boundary.

3. Currently due to the limitation of time, our downstream task relies on manual operation on images for landmark identification. In future, evaluation tasks can be designed through a comparison with real X-ray images.

4 What we've learned

1. How to manage computation resources: In this project, large datasets are included along with great amount of code development. We combined publicly available resources (e.g. Google Colab) with workstation of our lab, with nicely planned running and debugging of code.

2. How to work with large medical image: In this project, we learn the process for analysis medical images. Different from general images, coding with medical images need us to pay more attention to its formation, size and precision. And the method to evaluate the result is far more than directly comparison of some measurement.

3. Methods for evaluation in the absence of ground truth. Due to the lack of ground truth data in pair-wise X-ray image and CT images, we mannually design some artifitial data for this stage of evaluation.