Innovation/Impact: Deformable image registration(DIR) is widely used in radiation oncology despite lacking a ground truth for validating individual results. Quality assurance of DIR in the clinic relies on either visual inspection, or performance in phantom (benchmarking). However, benchmarking cannot be used to QA individual image registrations. It is essential to have a reliable quality assessment system to provide clinically meaningful evaluation. Here we develop a 3D neural network to automatically learn highly-representative features to quantify DIR errors as a first pass QA procedure. To deal with various DIR algorithms and differing anatomy, it is expected to have good robustness as well. Motivated by the above intentions, a 3D neural network, based on registered images labelled by expert identified point landmarks, has been demonstrated to have reliable quality inference. In this study, we evaluate and improve the deep learning based registration QA algorithm robustness by 1) Improving preprocessing and network design; 2) evaluating generalizability of the model with new registration algorithm DRAMMS.

Key Points: Figure 1 demonstrates basic data preparation procedure as the input of neural network, by taking an example of one patch. Figure 2 shows our neural network structure. Figure 3 visualized our quality assurance results mapped on original image:(a), overlapping of fixed and registered moving image;(b)-(d), predicted label map overlay; (e)-(g), true label map overlay, in which the highlighted areas have worse registration quality. Table 1 records the overall accuracy and Area under Curve(AUC) of our model, tested on various datasets and DIR algorithms. Figure 4 is the ROC curve of three models, all tested on Long4DCT / DRAMMS.

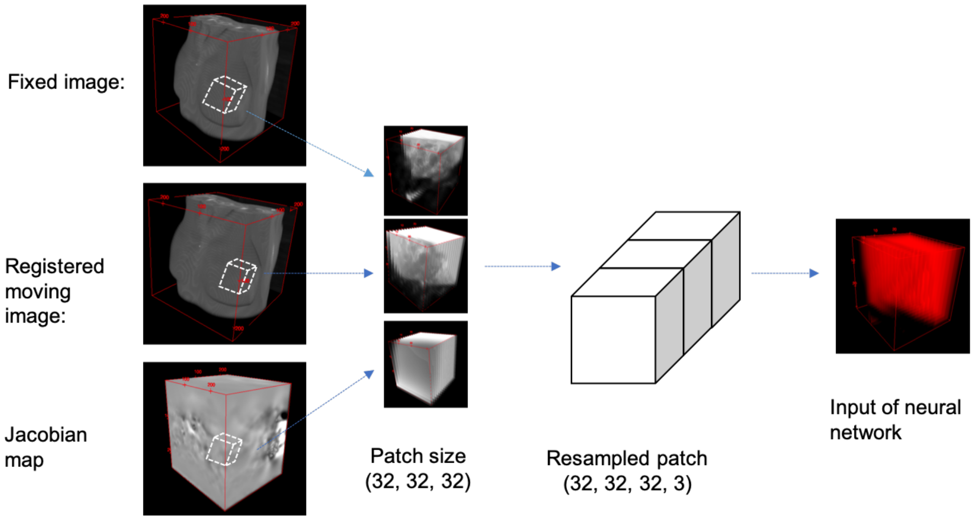

Figure 1: Image preparation procedure of one patch, as input of neural network

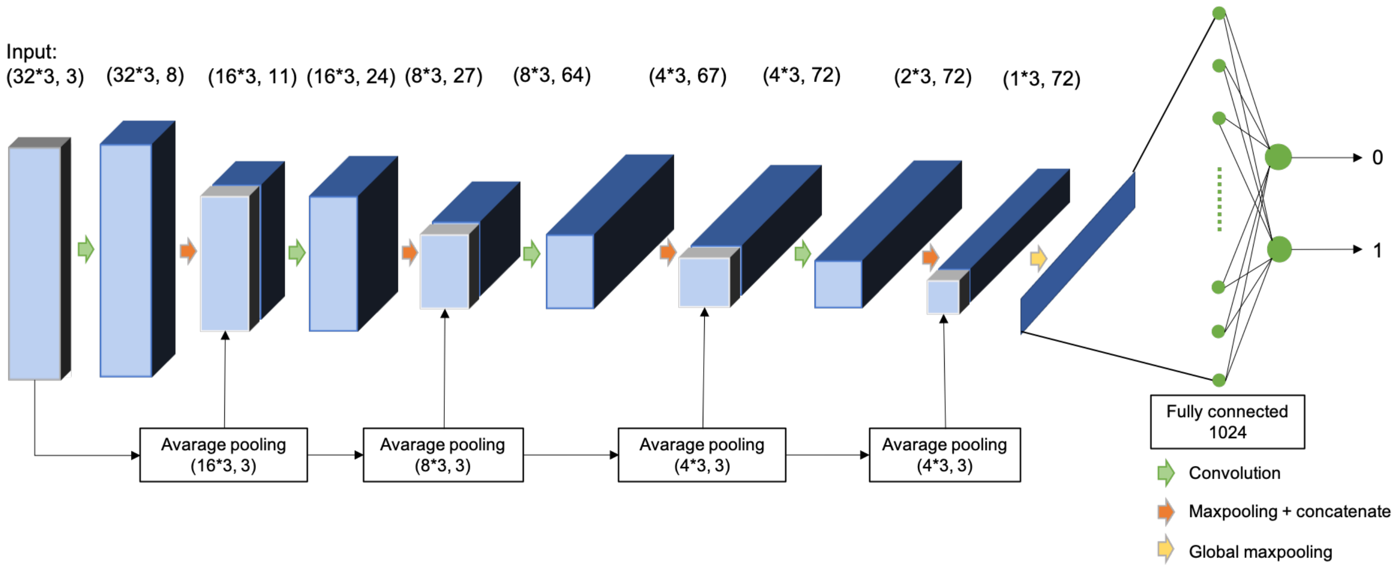

Figure 2: Structure of our neural network, input: resampled patches, labeled by its landmark distance error

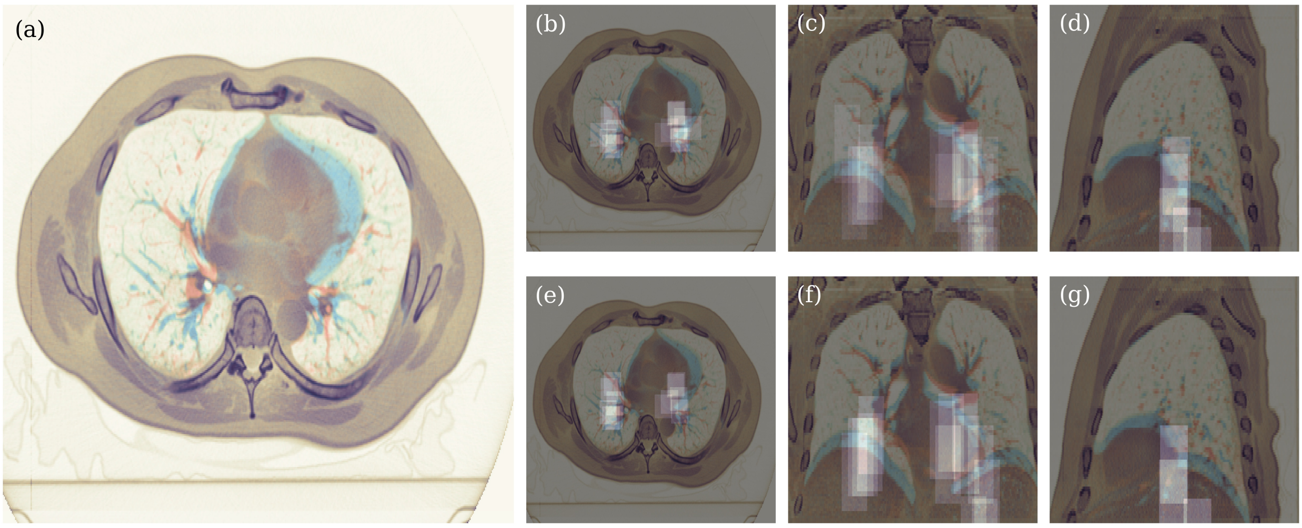

Figure 3: Visualized Quality Assurance results of our model, mapped with original image:(a), overlapping of fixed and registered moving image; (b)-(d), predicted label map overlay; (e)-(g), true label map overlay. The highlighted areas have worse registration quality.

Training		**Held-out Evaluation**				**Out-of-sample Evaluation**			
Data	**Regist-ration**	**Data**	**Regist-ration**	**Accuracy**	**AUC**	**Data**	**Regist-ration**	**Accuracy**	**AUC**
Dirlab	B-spline	Dirlab	B-spline	0.95	0.99	Long4DCT	B-spline	0.88	0.94
						Long4DCT	DRAMMS	0.87	0.94
						Dirlab	DRAMMS	0.89	0.96
Dirlab	B-spline	Dirlab	B-spline	0.95	0.99	Long4DCT	B-spline	0.88	0.94
Dirlab	DRAMMS	Dirlab	DRAMMS	0.95	0.99	Long4DCT	DRAMMS	0.87	0.93
Dirlab	B-spline	Dirlab	B-spline	0.95	0.99	Dirlab	DRAMMS	0.92	0.97
Long4DCT	B-spline	Long4DCT	B-spline	0.91	0.97	Long4DCT	DRAMMS	0.90	0.96

Table 1: **Overall Accuracy and Area under Curve(AUC) of our model**

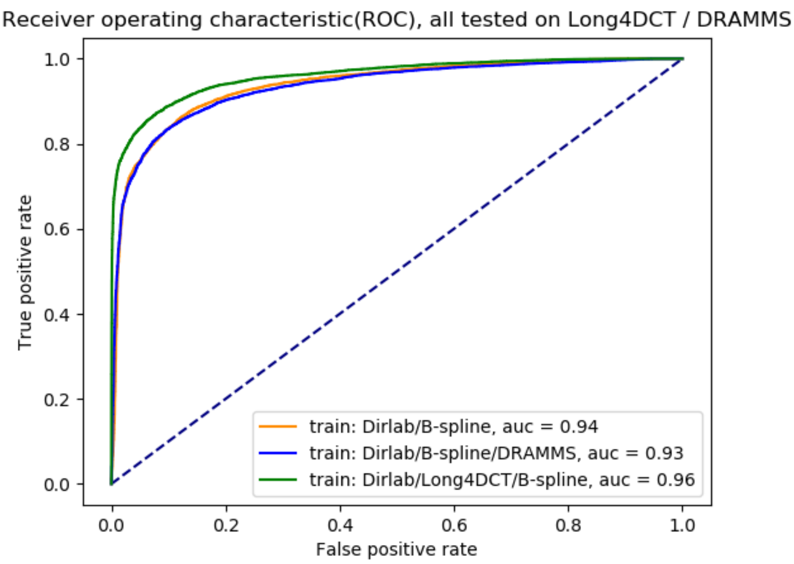

Figure 4: ROC curve of three models, trained by different datasets, tested on Long4DCT / DRAMMS