Image Reconstruction for Proton Therapy Range Verification via U-NETs

Setterdahl, Lena M.; Lionheart, William R.B.; Holman, Sean; Skjerdal, Kyrre; Ratliff, Hunter N.; Smeland Ytre-Hauge, Kristian; Lathouwers, Danny; Meric, Ilker

DOI
10.1007/978-3-031-66955-2_16

Publication date
2024

Document Version
Final published version

Published in
Medical Image Understanding and Analysis

Citation (APA)
Setterdahl, L. M., Lionheart, W. R. B., Holman, S., Skjerdal, K., Ratliff, H. N., Smeland Ytre-Hauge, K., Lathouwers, D., & Meric, I. (2024). Image Reconstruction for Proton Therapy Range Verification via U-NETs. In M. H. Yap, C. Kendrick, A. Behera, T. Cootes, & R. Zwiggelaar (Eds.), *Medical Image Understanding and Analysis: 28th Annual Conference, MIUA 2024 Manchester, UK, July 24–26, 2024 Proceedings, Part I* (pp. 232-244). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. LNCS 14859). Springer. https://doi.org/10.1007/978-3-031-66955-2_16

Image Reconstruction for Proton Therapy Range Verification via U-NETs

Lena M. Setterdahl[1(✉)], William R. B. Lionheart[2], Sean Holman[2], Kyrre Skjerdal[1], Hunter N. Ratliff[1], Kristian Smeland Ytre-Hauge[3], Danny Lathouwers[4], and Ilker Meric[1]

[1] Department of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen, Norway
{Lena.Marie.Setterdahl,Kyrre.Skjerdal,Hunter.Nathaniel.Ratliff,Ilker.Meric}@hvl.no

[2] Department of Mathematics, University of Manchester, Manchester, UK
{Bill.Lionheart,Sean.Holman}@manchester.ac.uk

[3] Department of Physics and Technology, University of Bergen, Bergen, Norway
Kristian.Ytre-Hauge@uib.no

[4] Department of Radiation Science and Technology, Delft University of Technology, Delft, Netherlands
D.Lathouwers@tudelft.nl

Abstract. This study aims to investigate the capability of U-Nets in improving image reconstruction accuracy for proton range verification within the framework of the NOVO (Next generation imaging for real-time dose verification enabling adaptive proton therapy) project. NOVO aims to enhance the accuracy of proton range verification by imaging the distribution of prompt gamma-rays (PGs) and fast neutrons (FNs) produced by nuclear interactions within tissue. In this work, focus lies on FNs, leaving PGs as future work. A dataset consisting of Monte Carlo-based simple back-projection and ground truth images of FN production distributions in a homogeneous water phantom was utilized. Various U-Net models were trained to predict FN distributions, and a set of range landmark (RL) metrics were computed for evaluation. Linear regression models were established to correlate shifts in mean RL with true range shift magnitudes. Our findings demonstrate a strong linear correlation between the shifts in mean RL in U-Net predictions and the true range shift magnitudes. Multiple RL metrics, including weighted average, inflection point, edge, and peak, were explored. This study highlights the potential utility of U-Nets in enhancing image reconstruction accuracy for proton range verification. The observed correlations between RL shifts and true range shifts provide evidence of the ability of U-Nets to accurately predict images containing range information. Future research will focus on generating more realistic training data incorporating more clinically relevant phantoms, including tissue heterogeneities.

Keywords: Proton therapy · Range verification · U-Net

M. H. Yap et al. (Eds.): MIUA 2024, LNCS 14859, pp. 232–244, 2024.
https://doi.org/10.1007/978-3-031-66955-2_16

1 Introduction

Proton therapy is a highly attractive radiotherapy treatment for cancer due to the sharp dose gradients and greater healthy tissue sparing that it offers as compared to conventional radiotherapy with photons [9]. However, its full potential is limited by range uncertainties caused by anatomical motion and day-to-day variations, tissue and tumor changes in response to treatment, patient setup or positioning errors, and inevitable uncertainties in the conversion of Computed Tomography (CT) Hounsfield units to relative proton stopping powers [5,13]. Consensus holds that range uncertainties, determined through a planning CT scan, are commonly found in the range of ± 3% [10]. To harness the full potential benefits of proton therapy, it is essential to accurately predict the range of proton beams during treatment planning and delivery. Inaccurate estimation of safety margins can lead to more significant repercussions in proton therapy compared to photon therapy. While underestimated margins in photon therapy may result in tumor under-dosage, in proton therapy, such underestimation could lead to portions of the tumor receiving no dose due to potential shifts in the sharp distal dose fall-off.

Numerous non-invasive range verification systems have been proposed to mitigate range uncertainty in proton therapy. These systems hinge on imaging the emission probability distribution of secondary particles resulting from proton interactions with tissue, including prompt gamma-rays (PGs), positron emitters, and fast neutrons (FNs), as these distributions exhibit strong correlation with the beam range. The most common method involves the imaging of PGs, with proposed systems such as prompt gamma-ray timing, prompt gamma-ray spectroscopy, and Compton Cameras [4,21], and imaging of positron emitters, where distributions thereof can be obtained by means like Positron Emission Tomography (PET) [3,7].

Achieving millimeter-level precision presents a notable challenge due to limited statistics for each proton beam spot, constraints related to hardware and readout electronics [2,15], and also inherent limitations within conventional image reconstruction algorithms - including Simple back-projection (SBP), Maximum Likelihood Expectation Maximization (MLEM), and Origin Ensembles (OE).

The iterative nature of MLEM and OE algorithms, although powerful, can produce images with excessive noise (MLEM) or blurred details (OE) [6], and SBP, while relatively simple and fast, may result in suboptimal image reconstructions, not being able to adequately address the challenges of measurement uncertainties, especially in scenarios with limited statistical data, which are common in proton therapy range verification.

U-Nets, a convolutional neural network architecture with a distinctive "U" shape, are becoming increasingly popular in various medical imaging modalities such as PET, Magnetic Resonance Imaging (MRI), and CT for enhancing the image reconstruction process. The design of U-Nets consists of encoding and decoding (down and up sampling/pooling) layers and skip connections between corresponding layers, facilitating the preservation of spatial information,

making it particularly effective for tasks where retaining spatial information is crucial, such as in medical imaging. Originally designed for MRI image segmentation [20], U-Nets have demonstrated versatility beyond their initial purpose and have found successful applications in tasks such as noise reduction and enhanced spatial resolution in PET [16] and prediction of dose distributions in radio therapy [11].

In this work we explore the potential of U-Nets for proton range verification in context of the NOVO (Next generation imaging for real-time dose verification enabling adaptive proton therapy) project, which aims to improve the accuracy of proton therapy range verification through imaging of the production distributions of both PGs and FNs. A core component of NOVO is the Compact Detector Array known as NOVCoDA [8], composed of optically segmented, densely stacked organic scintillator bars with light read-out at both ends. Imaging using the NOVCoDA relies on the assumption that the origin of detected particles, FNs and PGs, lie on the surface of a so-called event cone. Reconstructed event cones serve as input to the tomographic reconstruction of the particle production distributions used for range verification.

We consider the limited case of range shifts and production of FNs in a water phantom and train various U-Net models on Monte Carlo-generated SBP images and their respective ground truth (GT) FN production distribution. Central to our investigation is the assessment of whether the images predicted by U-Nets contain accurate range information. To evaluate the performance of the trained models, we compute a range landmark (RL) metric for the lateral profile (i.e., the profile parallel to the proton beam axis) of predicted images and establish a linear regression model to correlate shifts in mean RL with true range shift magnitudes. Performance of trained U-Nets are evaluated based on the coefficient of determination and slope of linear regression models. We explore the performance of eight different RL metrics computed from the lateral FN profiles, including weighted average, inflection point, edge, peak, and the 50% and 80% points of the edge and peak. Lastly, we discuss potential use cases for U-Nets within the NOVO image reconstruction system and outline avenues for future research.

2 Methods

2.1 U-Net

A U-Net architecture, comprising of encoding and decoding layers ($3 \cdot 10^7$ parameters), was implemented using the PyTorch library [14]. The encoder layers are responsible for extracting features from the input image through a series of convolutional and pooling operations. Each encoder layer captures increasingly abstract representations of the input image, starting from low-level features and progressively incorporating higher-level information. On the other hand, the decoder layers decode these abstract representations to reconstruct the output image. Each decoder layer involves up-sampling and convolutional operations, allowing the network to recover spatial details lost during encoding. In the forward pass, the input image undergoes encoding and decoding stages within the

U-Net architecture. Finally, the output image is normalized to ensure that pixel values fall within the range of [0, 1], facilitating consistency and compatibility with subsequent processing steps.

A collection of models underwent training through a grid search across various hyperparameters, encompassing two separate learning rate schedulers (ExponentialLR with $\gamma = \{0.7, 0.9\}$ and StepLR with step size 5 and $\gamma = 0.9$), two base learning rates (10^{-3} and 10^{-4}), and three distinct loss functions: Mean Square Error (MSE), L1, and Structural Similarity (SSIM) loss.

The MSE and L1 loss functions were directly obtained from the PyTorch library, offering traditional measures for assessing the difference between predicted and GT images. Meanwhile, the SSIM loss function was formulated as $SSIMLoss = 1 - SSIM$, where SSIM is calculated using the SSIM function provided by the `piqa` library [17], with a window size of 5. Each loss function offers distinct characteristics: MSE loss favors smoothness in the predicted images, while L1 loss tends to preserve sharpness in edges. On the other hand, $SSIMLoss$ is specifically designed to prioritize images with similar mean intensity, contrast, and structural information within local regions of a specified size, capturing important visual features, such as luminance, contrast, and structures (such as local patterns, textures and edges) present in the GT images [12,22].

Models were trained on SBP images as input and GT images as target for 50 epochs with a batch size of 10 and the Adam optimizer on an NVIDIA Tesla P4. The generation of SBP and GT images are explained in the subsequent section.

2.2 Image Generation: Monte Carlo Simulations, Bootstrapping and Image Pre-processing

Images for training the U-Net models were generated using Monte Carlo simulations with the GATE software (v.9.0) [18]. Placed in vacuum, a $40 \times 40 \times 40$ cm^3 water phantom was irradiated with an 85 MeV (57.8 mm range in water [1]) pencil proton beam featuring a $\sigma_x = \sigma_z = 2$ mm Gaussian beam profile and an intensity of 10^9 protons, equivalent to a high-intensity treatment beam spot. Positioned with a 10 cm gap between the phantom and detector surface, a $20 \times 30 \times 20$ cm^3 detector volume composed of typical scintillator plastic (with carbon-to-hydrogen atom ratio 10:11 and density 1.099 g/cm^3) was utilized. A set of 11 range shifts were introduced by incrementally adding or removing material from the phantom along the beam direction in 1 mm steps, emulating physiological changes like a patient gaining or losing weight in a region intersected by the proton beam. The setup is visualized in Fig. 1.

The physics lists QGSP_BIC_EMY was selected to focus on the relevant neutron energies, being the recommended physics list for proton therapy related simulations [8]. For this study, we focused solely on FN production within the water phantom. True FN production coordinates were used to generate GT images.

An artifact in the form of an unexpected peak in the FN production distributions was observed at the interface between vacuum and water phantom. This

artifact was excluded from the analysis to ensure accurate image reconstruction and interpretation of the results.

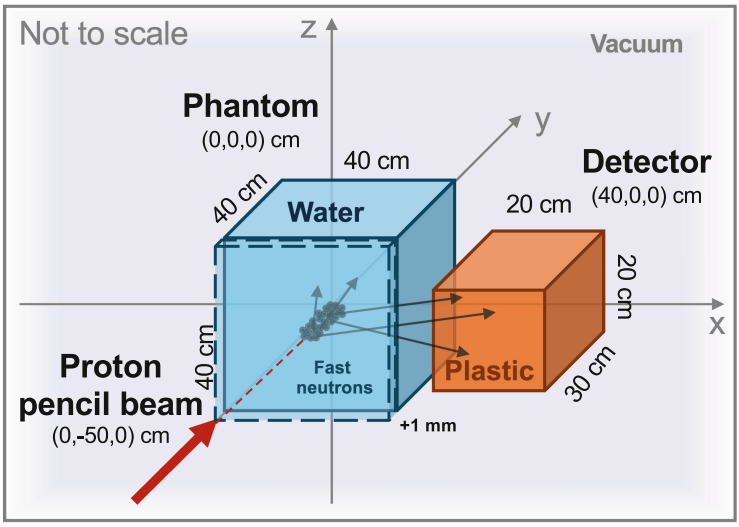

Fig. 1. GATE simulation geometry (not to scale) of an 85-MeV pencil proton beam aimed perpendicular to a $40 \times 40 \times 40$ cm^3 water phantom and a $20 \times 30 \times 20$ cm^3 plastic scintillator detector volume placed 10 cm from phantom surface.

FNs undergoing two elastic scatters on hydrogen atoms in the detector volume, with an energy deposition greater than 0 eV in each, were used for event-cone reconstruction, using non-relativistic scatter kinematics to compute the half opening angle, as in [8]. Detector resolution effects were not accounted for. Using a similar approach as the marching algorithm [23], SBP images were constructed by projecting event cones back onto an image plane positioned on and with a surface normal perpendicular to the proton beam axis. Event cones pointing in the opposite direction of the phantom were filtered out.

To augment the training data, 200 bootstrap operations were performed on each range shift simulation, sampling n event cones and respective GT coordinates of the observed FN double coincident scatter events, yielding a total of 2200 pairs of SBP-GT images. The number of bootstrap samples for a desired beam intensity N was determined by

$$n = \eta N, \tag{1}$$

where the FN double scatter efficiency η is calculated as number of FN double scatter per primary proton, $\eta = (1.86 \pm 0.11) \cdot 10^{-4}$ FN double scattering events per primary proton. To emulate a medium intensity proton beam spot, we set $N = 10^8$, resulting in $n = (1.86 \pm 0.11) \cdot 10^4$ FN double scattering events.

The SBP and GT images were cropped to dimensions of 64×64 pixels with each pixel measuring 1 mm $\times$ 1 mm and normalized such that maximum intensity in a given image pixel equaled 1. SBP-GT image pairs were utilized for training, validation, and testing, with SBP images serving as model inputs and GT images as targets. The dataset was split into training, validation, and test subsets using an approximate ratio of 60-20-20%, resulting in 1320, 429, and 451 image pairs for each respective subset, leaving 40 image pairs of each range shift for the test subset.

2.3 Range Landmark and Performance Metric

Central to our study is the ability of trained U-Nets to accurately predict images that contain range information. This capability is assessed through the use of a so-called range landmark (RL) to quantitatively measure shifts in the lateral profile of reconstructed FN distributions.

In previous work [19], "area under the curve" (AUC) of 1D FN distribution profiles was deemed the most predicative feature, amongst 28 distribution parameters, for range shift detection. This conclusion was, however, based on simulations of FN production in a CT-based phantom where production profile of FNs were observed to take different shapes depending on the magnitude of the introduced range shift. Assuming that the observed change in FN distribution shape was due to the heterogeneous nature of the phantom, drastic changes in the shape and total number of produced FN are not expected in this work, where range shift deviations were introduced to a homogeneous water phantom, thus making AUC an inapt RL metric to determine range shifts.

Instead, we consider weighted average (as in [8]), inflection point, peak, edge, and the 50% and 80% points of both peak and edge as RL metrics (illustrated in Fig. 2), as these will be affected by shifts in FN distributions regardless of the distribution shape.

Peak RL was defined as the location of the lateral profile maximum. An algorithm was developed to find the edge of the lateral profile: The algorithm iterates through the gradient values of a lateral profile within a sliding window and computes the average gradient. The index of the starting bin of the corresponding window is returned as the edge value if the average gradient is the smallest among the considered windows. The sliding window size was defined as one-eights of the total number of bins in the lateral profile. Prior to computing RL metrics, the images underwent a 2x up-sampling process to create a finer binning (1 bin = 0.5 mm). The up-sampling process employed first-order spline interpolation between pixels and was facilitated by the `zoom` function available in the `scipy.ndimage` module.

To gauge performance, U-Net models were evaluated by fitting a linear regression model to determine the correlation between shifts in mean RL ($\Delta\bar{RL}$) and true range shift magnitudes. $\bar{RL}$ was determined by averaging over a set of 40 RL for corresponding true range shift, where 40 is the number of bootstraps per true range shift in the test subset as mentioned in Sect. 2.2. The shift in average RL was found by taking the difference between a $\bar{RL}$ and the $\bar{RL}$ for 0 mm

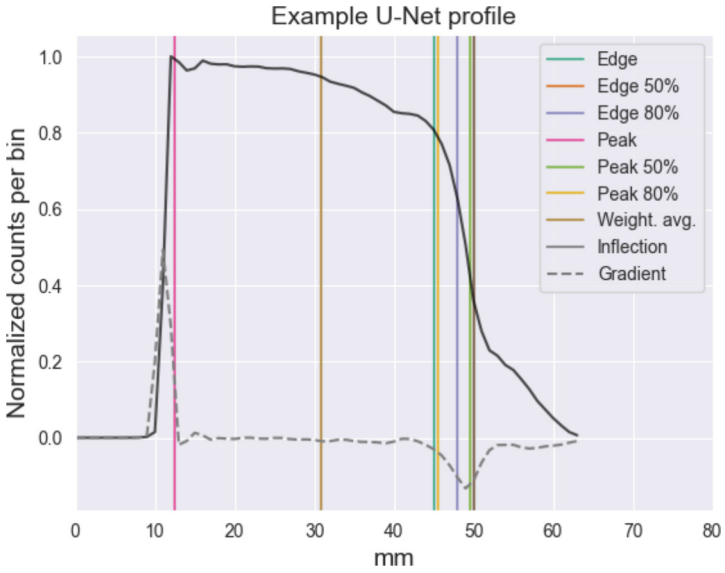

Fig. 2. Example of the lateral profile (black) and its first derivative (gradient) of a U-Net predicted fast neutron distribution and corresponding range landmarks indicated on the figure.

true range shift for the RL metric under consideration, and the corresponding standard deviation was set as the standard deviation of the RL set in question

$$\sigma_{\Delta\bar{RL}} = \sigma_{\bar{RL}} = \sqrt{\sum_{k}^{40} \frac{(RL_k - \bar{RL})^2}{40}}. \tag{2}$$

When examining the parameters of the linear regression model, it is desirable to have a coefficient of determination R^2 and a slope a that are close to 1, both features equally desirable. A coefficient of determination of $R^2 = 1$ indicates a perfect fit, while a slope of $a = 1$ implies that for every unit increase in the true range shift, the $\Delta\bar{RL}$ of U-Net predicted FN distribution increases by exactly 1 unit, indicating a strong linear correlation.

The presence of a high R^2 does not guarantee an a close to 1, and vice versa. Meaning, a possible outcome of the linear regression analysis, could for instance, be a high R^2 and slopes close to zero, which would mean a good linear fit but a weak linear correlation. For this reason, models where $a < 0.1$ were removed from further evaluation, regardless of the R^2 value. No criteria for R^2 were enforced. Linear regression analysis was done for all combinations of trained U-Nets and eight RL metrics. Linear regression results were then grouped into RL metrics and was sorted according to descending R^2 and a.

Furthermore, the mean and standard deviations of the linear regression results (R^2 and a) were calculated for each RL metric to evaluate the impact of

the chosen RL metric on linear regression results. Also, RL metrics were compared by assessing the magnitude of the average RL standard deviation ($\bar{\sigma}_{\Delta\bar{R}L}$) of each metric, since it is crucial to have a metric that precisely identifies RL shifts.

3 Results and Discussion

Average and standard deviation of linear regression results (coefficient of determination R^2 and slope a) and $\sigma_{\Delta\bar{R}L}$ for each RL metric are reported in Table 1. Among the considered RL metrics, weighted average resulted in the lowest average $\sigma_{\Delta\bar{R}L}$, with a value of 1.46 ± 0.51 mm, and linear regression models with the highest average and lowest standard deviation of R^2. The weighted average method appears to outperform the other RL metrics in range estimation, likely due to the peak and edge based RL metrics being more sensitive to statistical fluctuations in the lateral profile. For instance, inadequate window size in the edge RL detection algorithm may lead to identification of a local minima instead of the edge of a profile. It should be highlighted that RLs were computed based on images predicted by U-Net models and that the performance of the RL weighted average approach cannot be extrapolated to FN distribution images reconstructed by other means. Moreover, it is important to clarify that this study does not include an analysis to determine the range shift detection limits of U-Nets, leaving this aspect for future investigation.

Table 1. Average and standard deviation of coefficient of determination R^2 and slope a of linear regression fit, and shifts in mean range landmark $\Delta\bar{R}L$ for different range landmark metrics.

Range landmark	$\bar{R}^2$	σ_{R^2}	$\bar{a}$	σ_a	$\bar{\sigma}_{\Delta\bar{R}L}$ [mm]	$\sigma_{\bar{\sigma}_{\Delta\bar{R}L}}$ [mm]
Weighted average	0.97	0.02	0.80	0.11	1.46	0.51
50% of peak	0.93	0.08	0.93	0.18	2.48	1.63
80% of peak	0.91	0.06	1.06	0.21	3.31	1.79
50% of edge	0.91	0.10	0.87	0.13	4.16	5.23
Edge	0.91	0.10	0.88	0.12	4.70	4.75
80% of edge	0.91	0.10	0.88	0.12	4.70	4.75
Inflection point	0.88	0.11	1.03	0.26	3.40	2.49
Peak	0.85	0.07	0.93	0.17	3.85	1.87

The top three U-Net models according to linear regression for RL computed by weighted average are shown in Fig. 3, and corresponding linear regression results in addition to average and standard deviation of $\sigma_{\Delta\bar{R}L}$ are listed in Table 2. The primary aim of this investigation is to evaluate the effectiveness of U-Nets in predicting images containing range shift information, a task for which their remarkable coefficient of determination and slope demonstrate their

capability, all top three models displaying an $R^2 > 0.98$ and $a > 0.8$ (see Table 2). A large fraction of the models exhibited notable performance, regardless of the combination of loss function, learning rate scheduler and parameter γ, and base learning rate employed. However, a trend emerges favoring a lower base learning rate of 10^{-3} and an exponential learning rate schedule with $\gamma = 0.9$.

Table 2. Top three U-Net models according to linear regression picked for highest coefficient of determination R^2 and slope a for weighted average as range landmark metric.

Model ID	Loss function	Scheduler	Base lr	γ	R^2	a	σ_a	b [mm]	σ_b [mm]	$\bar{\sigma}_{\Delta\bar{RL}}$ [mm]	$\sigma_{\sigma_{\Delta\bar{RL}}}$ [mm]
0	L1	Exp	10^{-3}	0.9	0.991	0.868	0.089	−0.418	0.028	1.180	0.401
1	MSE	Step	10^{-3}	0.9	0.988	0.899	0.104	−0.625	0.033	1.110	0.343
2	SSIM	Exp	10^{-3}	0.9	0.983	0.849	0.116	−0.274	0.037	1.150	0.361

Abbreviations: Exponential learning rate scheduler (Exp.), Step learning rate scheduler (Step), Base learning rate (Base lr)

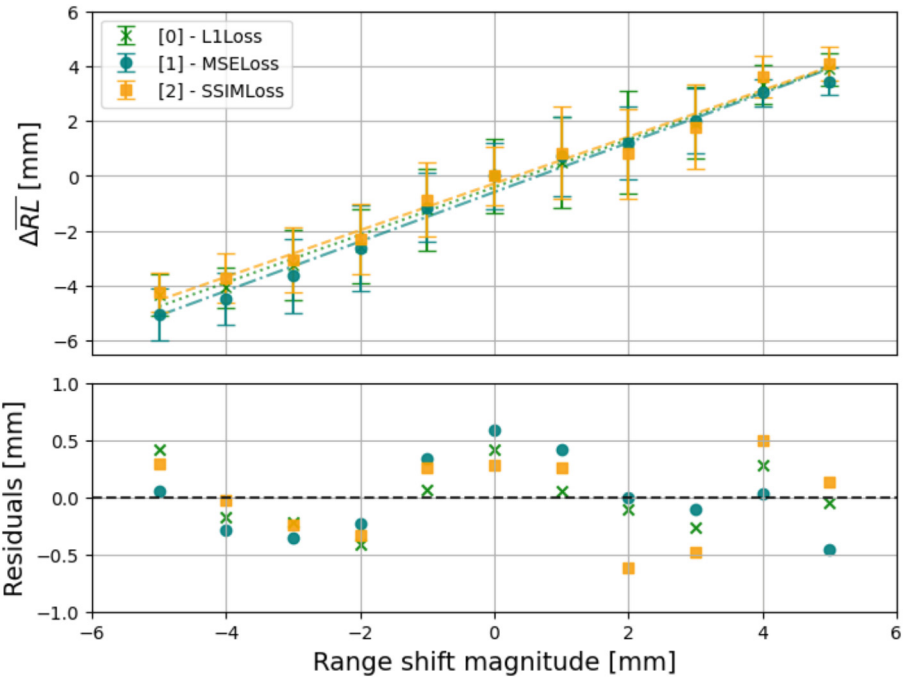

Fig. 3. Linear regressions of predicted image average range landmark (weighted average) shift and true range shift magnitude for the top three trained models with the labels corresponding to the model IDs in Table 2.

We present in Fig. 4 the predicted FN distributions from the top three models, each of which were trained on a different loss function, along with their respective input SBP and target GT images for 5 range shifts (−5, −3, 0, 3, and 5 mm). These exemplary images showcase the influence of the chosen loss function on model prediction. Specifically, the model trained with an SSIM loss function exhibits a preference towards preserving structure, mirroring patterns of the input SBP. Conversely, models trained with L1 and MSE losses produce smoother lateral profiles, exemplified in Fig. 5 which depicts the average lateral profiles of 40 images predicted by the models for the specific case of a 5 mm true range shift.

These analyses underscore the key role of the loss function in shaping the predicted FN distribution, an aspect not readily apparent in the linear regression analyses. Hence, in future endeavors employing U-Nets for FN distribution prediction in proton therapy, we recommend careful consideration of the loss function and the intended application of the predicted image. For instance, the structural fidelity of predicted distributions may be of immediate concern if the whole image (and not just the lateral profile) is to be used to compute a RL metric. Consequently, employing an SSIM loss function that prioritizes local structures and contrast levels may be more advantageous in certain scenarios compared to an L1 loss function that emphasizes sharp edges or an MSE loss function that emphasizes smoother edges.

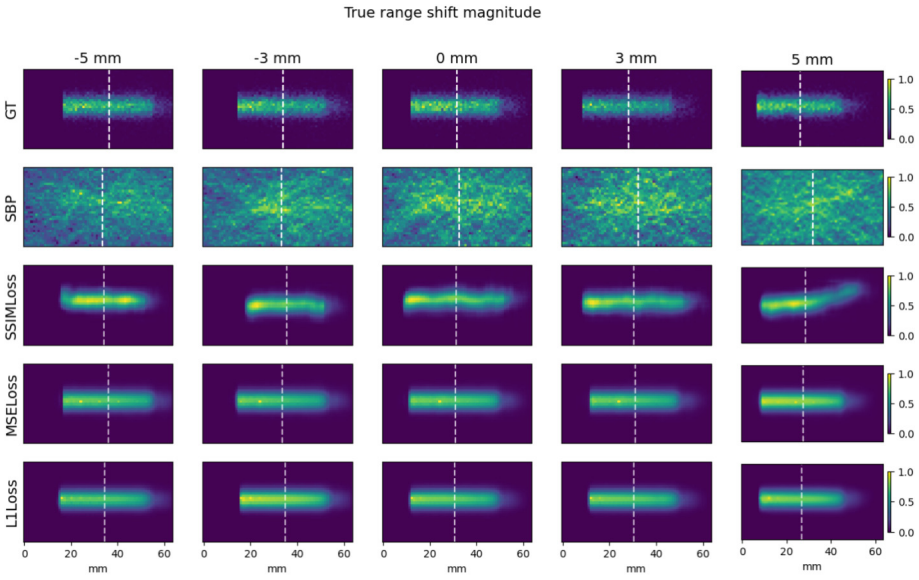

Fig. 4. Ground truth (GT), simple back-projections (SBP), and U-Nets (model 0, 1, and 2, trained with L1, MSE, SSIM loss, respectively) predictions of fast neutron production distribution and corresponding range landmark (weighted average) indicated by a dotted vertical line. All images are normalized such that maximum intensity is equal to 1.

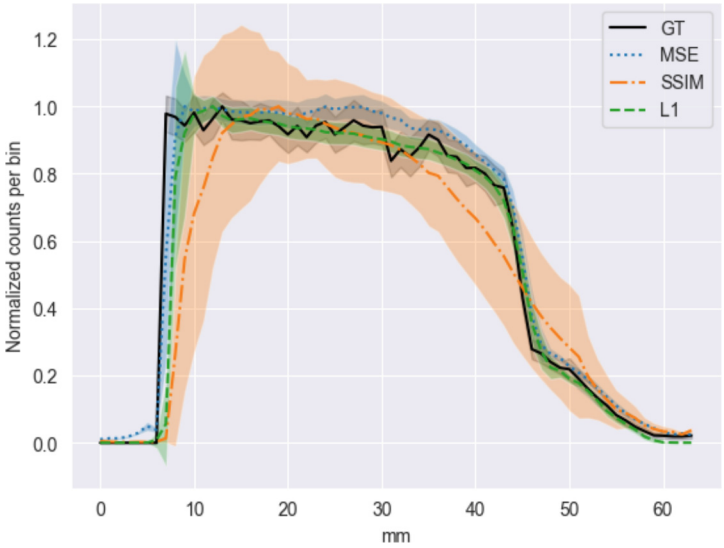

Fig. 5. The average lateral profiles of 40 images predicted by model 0, 1, and 2 (trained on L1, MSE, and SSIM loss function, respectively) for the specific case of a 5 mm true range shift.

This work's main priority has been to investigate the use of machine learning to enhance image reconstruction in proton therapy range verification. While a U-Net architecture was chosen for this purpose, other approaches, such as Generative Adversarial Networks (GANs), may also be viable options for future research.

4 Conclusion

In this study, we investigated the potential of U-Nets for proton therapy range verification within the framework of the NOVO (Next generation imaging for real-time dose verification enabling adaptive proton therapy) project by training models to predict fast neutron distributions based on simple back-projection images. A strong linear correlation between the shift in mean range landmark (RL) of U-Net predictions and the true range shift magnitude was observed, suggesting that U-Net models have the capability to predict images containing range information. A significant portion of trained models exhibited strong linear correlations. While these findings are promising and underscore the potential utility of U-Nets in improving image reconstruction accuracy for proton range verification, we emphasize the limitations of this study and the need for a more diverse dataset to provide more conclusive evidence. In future work, we suggest generating realistic training data, considering phantom heterogeneities and various clinical proton beam energies, intensities, and directions. In conclusion, while our study provides valuable insights into the potential of U-Nets in image

reconstruction for the NOVO project, there remains a need for ongoing research and development to fully harness their capabilities.

Acknowledgments. The authors would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Rich Non-linear Tomography, where work on this paper was undertaken. This work was supported by EPSRC (Grant no.: EP/R014604/1 and EP/V007742/1), the Research Council of Norway (Grant no.: 301459), and the European Innovation Council (Grant no. 101130979).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Berger, M., Coursey, J., Zucker, M.: ESTAR, PSTAR, and ASTAR: computer programs for calculating stopping-power and range tables for electrons, protons, and helium ions (version 1.21) (1999). http://physics.nist.gov/Star
2. Golnik, C., et al.: Tests of a Compton imaging prototype in a monoenergetic 4.44 mev photon field-a benchmark setup for prompt gamma-ray imaging devices. J. Instrum. **11**, P06009 (2016). https://doi.org/10.1088/1748-0221/11/06/P06009
3. Helmbrecht, S., Santiago, A., Enghardt, W., Kuess, P., Fiedler, F.: On the feasibility of automatic detection of range deviations from in-beam pet data. Phys. Med. Bio. **57**(5), 1387 (2012). https://doi.org/10.1088/0031-9155/57/5/1387
4. Hueso-González, F., Rabe, M., Ruggieri, T.A., Bortfeld, T., Verburg, J.M.: A full-scale clinical prototype for proton range verification using prompt gamma-ray spectroscopy. Phys. Med. Bio. **63**(18), 185019 (2018). https://doi.org/10.1088/1361-6560/aad513
5. Knopf, A.C., Lomax, A.: In vivo proton range verification: a review. Phys. Med. Bio. **58**(15), R131 (2013). https://doi.org/10.1088/0031-9155/58/15/R131
6. Kohlhase, N., et al.: Capability of MLEM and OE to detect range shifts with a Compton camera in particle therapy. IEEE Trans. Radiat. Plasma Med. Sci. **4**(2), 233–242 (2020). https://doi.org/10.1109/TRPMS.2019.2937675
7. Lu, H.M.: A potential method for in vivo range verification in proton therapy treatment. Phys. Med. Bio. **53**(5), 1413 (2008). https://doi.org/10.1088/0031-9155/53/5/016
8. Meric, I., et al.: A hybrid multi-particle approach to range assessment-based treatment verification in particle therapy. Sci. Rep. **13**(1), 6709 (2023). https://doi.org/10.1038/s41598-023-33777-w
9. Mohan, R.: A review of proton therapy - current status and future directions. Precis. Radiat. Oncol. **6**(2), 164–176 (2022). https://doi.org/10.1002/pro6.1149
10. Moyers, M., Miller, D.W., Bush, B.A., Slater, J.D.: Methodologies and tools for proton beam design for lung tumors. Int. J. Radiat. Oncol. Biol. Phys. **49**, 1429–38 (2001). https://doi.org/10.1016/s0360-3016(00)01555-8
11. Nguyen, D., et al.: A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. Sci. Rep. **9**(1), 1076 (2019). https://doi.org/10.1038/s41598-018-37741-x
12. Nilsson, J., Akenine-Möller, T.: Understanding SSIM (2020). https://arxiv.org/abs/2006.13846

13. Paganetti, H.: Range uncertainties in proton therapy and the role of monte Carlo simulations. Phys. Med. Bio. **57**(11), R99-117 (2012). https://doi.org/10.1088/0031-9155/57/11/R99
14. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library (2019). https://arXiv.org/abs/1912.01703
15. Pausch, G., et al.: Detection systems for range monitoring in proton therapy: Needs and challenges. Nucl. Instrum. Methods Phys. Res. A **954**, 161227 (2020). https://doi.org/10.1016/j.nima.2018.09.062, symposium on Radiation Measurements and Applications XVII
16. Reader, A.J., Pan, B.: Ai for pet image reconstruction. Brit. J. Radiol. **96**(1150), 20230292 (2023). https://doi.org/10.1259/bjr.20230292
17. Rozet, F.: PIQA: PyTorch image quality assessment.https://doi.org/10.5281/zenodo.7821598, https://pypi.org/project/piqa
18. Sarrut, D., et al.: A review of the use and potential of the gate monte Carlo simulation code for radiation therapy and dosimetry applications. Med. Phys. **41**, 064301 (2014). https://doi.org/10.1118/1.4871617
19. Schellhammer, S.M., Meric, I., Löck, S., Kögler, T.: Hybrid treatment verification based on prompt gamma rays and fast neutrons: multivariate modelling for proton range determination. Front. Phys. **11**, 1295157 (2023). https://doi.org/10.3389/fphy.2023.1295157
20. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(4), 640–651 (2016). https://doi.org/10.1109/TPAMI.2016.2572683
21. Verburg, J.M., Seco, J.: Proton range verification through prompt gamma-ray spectroscopy. Phys. Med. Bio. **59**(23), 7089–7106 (2014). https://doi.org/10.1088/0031-9155/59/23/7089
22. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
23. Wilderman, S., Rogers, W., Knoll, G., Engdahl, J.: Fast algorithm for list mode back-projection of Compton scatter camera data. IEEE Trans. Nucl. Sci. **45**(3), 957–962 (1998). https://doi.org/10.1109/23.682685