Proceedings of ICMMGH 2027 Symposium: Clinical OMICS for Drug Discovery and Personalized Medicine
DOI: 10.54254/2753-8818/2027.MB37493

Per-Structure ROI-Weighted Loss Improves DVH Accuracy in Deep Learning-Based Radiotherapy Dose Prediction on the OpenKBP Benchmark

Chengkun Ma

Department of Physics & Astronomy, University of Waterloo, Waterloo, Canada
c242ma@uwaterloo.ca

Abstract. Radiotherapy dose prediction networks usually train with a single, spatially uniform loss such as voxel-wise mean absolute error (MAE), which weights errors in small, clinically critical structures the same as errors in unremarkable tissue. This study uses the official OpenKBP benchmark to test whether adding a per-structure, region-of-interest (ROI)-weighted term to a standard masked-MAE loss improves dose-volume histogram (DVH) accuracy without degrading dose accuracy. A 3-D U-Net dose predictor was reimplemented in PyTorch on the official OpenKBP data split (200/40/100 patients). Nine training runs varied loss (MAE baseline, MAE + ROI-weighted, ROI-only ablation), filter count, λ, and seed. The primary analysis compared the baseline and proposed losses at the main model capacity on the held-out 100-patient test set with a two-sided Wilcoxon signed-rank test and the Hodges–Lehmann estimator, under a prespecified plan. The proposed loss reduced dvh_error for 99/100 test patients (mean 3.58 → 2.44 Gy; Hodges–Lehmann −1.13 Gy, 95% CI [−1.23, −1.04]; $p = 4.5\times10^{-18}$), supporting the primary hypothesis. dose_error also decreased (3.57 → 3.36 Gy) rather than being traded off. The effect held in direction under an independent seed and a smaller capacity, and an ablation confirmed the global MAE term remains necessary: removing it left dose_error far worse while leaving dvh_error unimproved. A simple, architecture-agnostic per-structure ROI-weighted term, added to rather than substituted for a standard masked-MAE loss, produced a robust, directionally consistent improvement in DVH accuracy on the OpenKBP benchmark without an offsetting dose-accuracy cost.

Keywords: radiotherapy dose prediction, deep learning, U-Net, dose-volume histogram, loss function design

1. Introduction

The goal of radiotherapy treatment planning is to deliver a prescribed dose to the tumor targets and to spare the nearby organs at risk (OARs). Conventional inverse planning is still an iterative, planner-dependent process that takes a lot of manual effort and expertise. Knowledge-based planning (KBP) uses information from prior plans to predict the achievable dose-volume metrics, dose-volume histograms (DVHs), or the spatial dose distributions for a new patient in order to

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2027.MB37493

support plan optimization, efficiency, consistency, and quality assessment [1]. The encoder–decoder networks derived from the U-Net [2, 3], first developed for 2-D and 3-D biomedical image segmentation, have since been widely used for voxel-wise dose prediction from CT images and structure contours [1].

The AAPM OpenKBP Grand Challenge set up a public, standard benchmark for this problem: data from 340 patients with head-and-neck cancer, an official 200/40/100 train/validation/test partition, standardized reference dose distributions, and two official metrics — dose_score, the voxel-wise error over the possible-dose region, and dvh_score, the errors in a specified set of structure-specific DVH metrics [4]. The OpenKBP baseline is trained with the voxel-wise mean absolute error (MAE) computed over the entire predicted volume, not restricted to the possible-dose region [4]. This is the uniform objective, which assigns the same error weight to every voxel, no matter whether it is in a target, in an OAR, outside the possible-dose region, or in other tissue.

This uniform weighting does not capture the structure-specific form of DVH evaluation: errors are looked at in the individual targets and OARs, while the masked global MAE used in this study looks at the errors over all the voxels in the possible-dose region. A small structure can therefore add little to the global loss even if its structure-specific dose metric is not accurate. In recent studies, domain information has been incorporated into dose-prediction objectives by using per-structure dose moments [5], a composite objective that contains structure-aware DVH-guided and spatial-gradient terms [6], and fixed structure-specific region-of-interest (ROI) weights [7]. Jhanwar et al. reported the improvement of DVH score with an MAE-plus-moment objective [5]; Mak et al. and Ogawa et al. show other structure- and distribution-aware loss designs [6, 7]. Other recent work has instead targeted the training strategy itself: a patient-specific adaptive-radiotherapy dose predictor, trained on only a single patient's own early-treatment data, outperformed a population-trained model [8]. The present study keeps the standard population-trained setup and instead modifies the loss function. These motivate the present question: when a structure-normalized ROI term is added to, rather than substituted for, the global objective, can DVH accuracy improve without the accompanying deterioration in whole-field dose accuracy?

In this study, a 3-D U-Net dose predictor was reimplemented in PyTorch on the official OpenKBP split to ask one prespecified primary question: does adding a per-structure, possible-dose-mask-intersected ROI-weighted term to a standard masked-MAE loss reduce dvh_error on the held-out test set? The dose_error metric was evaluated descriptively to see whether the observed DVH improvement was accompanied by worse overall dose accuracy. The ROI term is deliberately simple: it uses no structure moments, learned weighting, or architecture change, and it is normalized per patient and per structure so that neither a patient's field-of-view size nor a structure's voxel count can dominate it. The primary comparison used a two-sided Wilcoxon signed-rank test and the Hodges–Lehmann estimator under a plan fixed before the test set was evaluated. The four secondary or exploratory analyses assessed seed reproducibility, capacity robustness, loss ablation, and lambda sensitivity. The hypotheses, decision rules, and test-set evaluation conditions were fixed before test-set evaluation, and a code-version guard made sure that the analysis used the frozen training implementation. This design separates the prespecified primary comparison from the secondary and exploratory robustness and ablation analyses. Section 2 describes the data, network, loss, and statistical plan; Section 3 reports the results; Sections 4 and 5 discuss their interpretation, limitations, and conclusions.

DOI: 10.54254/2753-8818/2027.MB37493

2. Methods

A PyTorch 3-D U-Net was also trained, based on the official OpenKBP baseline [4], with either possible-dose-mask MAE or the proposed per-structure ROI-weighted objective, with the architecture, inputs, optimizer, and schedule fixed. Nine runs varied in filter count, loss, λ, and seed. The data, model, training, analyses, and the reproducibility safeguards are in Sections 2.1–2.7.

2.1. Dataset and official data split

OpenKBP contains 340 head-and-neck cases split into 200 training, 40 validation, and 100 test cases [4]. Each provides a planning CT, masks for PTV56/63/70 and seven OARs, and a synthetic reference dose generated by the dataset's automated planning pipeline rather than the original clinical dose [4]. Because test labels are now public, this partition is held out but not blind.

All nine runs were evaluated on the same 40 validation cases. The 100-case test set was reserved for H1–H3 and evaluated only by the four filters = 32 runs. H1 was primary; H2–H3 were descriptive/exploratory. The official loader returns the same ordered validation cases for every seed.

2.2. Input representation and preprocessing

CT and ten binary structure masks formed an 11-channel, 128^3 input; the output was one 128^3 dose channel. CT was divided by 1000 without clipping. Dose, loss, and score values remained in Gy; the possible-dose mask was binary. No augmentation was used.

2.3. Network architecture

Figure 1 shows the six-level 3-D U-Net, which follows the encoder–decoder and skip-connection principles of U-Net and 3D U-Net and the official OpenKBP topology [2-4]. Initial filters were f = 32 for the main model (~29.26 M parameters) and f = 16 for H4. Encoder blocks used stride-2 convolutions, BatchNorm3d, and LeakyReLU(0.2); the bottleneck omitted BatchNorm. Decoder blocks used transposed convolutions and LeakyReLU(0.0), with Dropout3d(0.2) only at dec4 and dec2. Five skip connections linked matching resolutions, including the final connection before the output head. PyTorch BatchNorm3d(momentum=0.01) matches Keras BatchNormalization(momentum=0.99) under their opposite conventions. All runs used Adam (β_1 = 0.5, β_2 = 0.999; learning rate 2×10^{-4}), batch size 2, and 100 epochs; final weights were used without best-checkpoint selection [9].

DOI: 10.54254/2753-8818/2027.MB37493

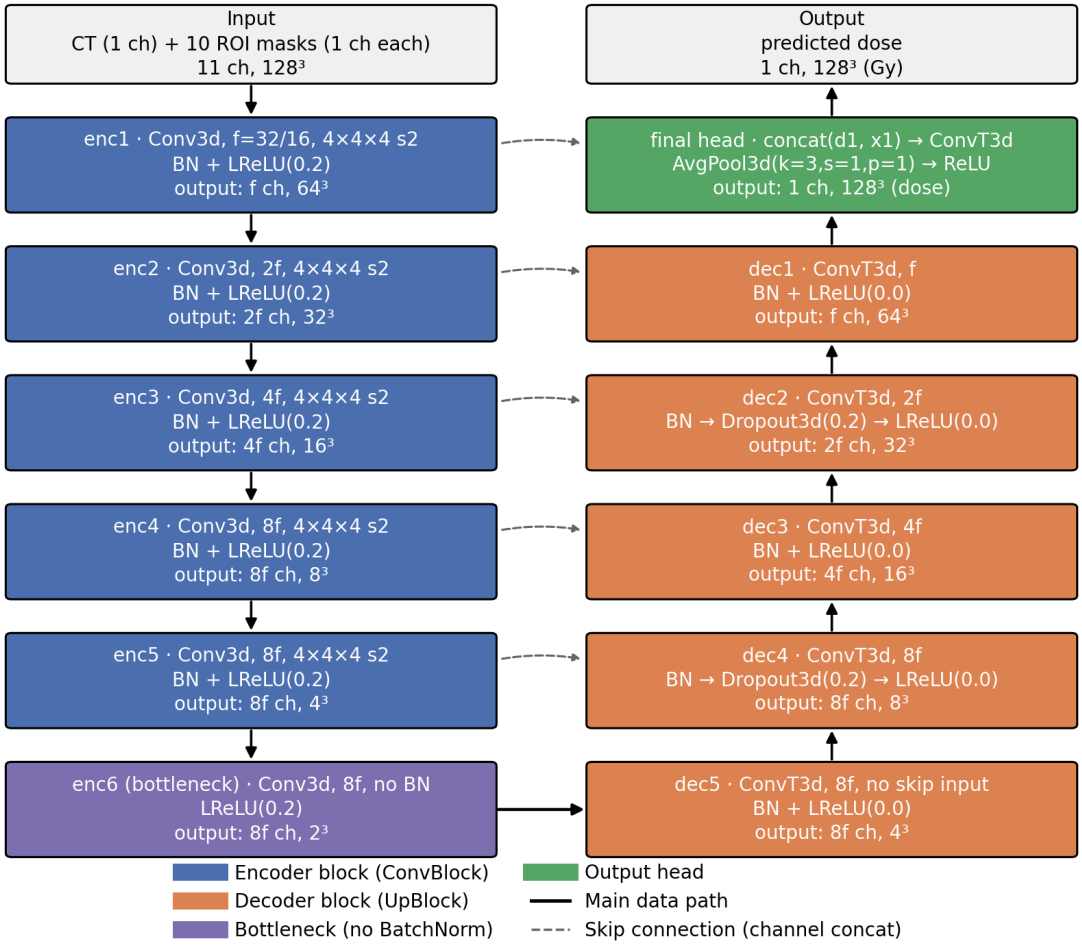

Figure 1. DoseUNet architecture. Six encoder stages (blue), a no-BatchNorm bottleneck (purple), five decoder stages (orange; Dropout3d at two stages), five dashed skip connections, and the output head (green). f = 32 for the main model and 16 for H4 (picture credit: original)

2.4. Loss functions

Global MAE was calculated within the possible-dose region, normalized per patient, and then averaged across the batch:

$$L_{\text{global}}^{(i)} = \frac{\sum_v M_{i,v} \left|\text{pred}_{i,v} - \text{target}_{i,v}\right|}{\sum_v M_{i,v}} \tag{1}$$

$$L_{\text{global}} = \left(\frac{1}{B}\right) \sum_{i=1}^{B} L_{\text{global}}^{(i)} \tag{2}$$

where i indexes patients, v voxels, B batch size, and M the possible-dose mask. Prior dose-prediction objectives used moment-based DVH surrogates [5], structure-aware DVH and spatial-gradient terms [6], or ROI-weighted MSE and dose-gradient terms [7]. The present objective instead averages MAE within each available ROI and then across structures and patients:

DOI: 10.54254/2753-8818/2027.MB37493

$$L_{ROI}^{(i)} = \left(\frac{1}{K_i}\right) \sum_{k:\,|R'_{i,k}|>0} \left[\left(\frac{1}{|R'_{i,k}|}\right) \sum_{v \in R'_{i,k}} \left|pred_{i,v} - target_{i,v}\right|\right] \tag{3}$$

$$L_{ROI} = \left(\frac{1}{B}\right) \sum_{i=1}^{B} L_{ROI}^{(i)} \tag{4}$$

Here, $R'_{i,k} = R_{i,k} \cap M_i$, and K_i is the number of non-empty ROI structures available for patient i after intersection with the possible-dose mask. Absent structures are excluded from both the summation and K_i . Per-ROI normalization prevents large structures from dominating small ones, and overlapping PTV voxels contribute once per PTV channel. Three loss conditions were tested baseline (mae):

$$L = L_{global} \tag{5}$$

proposed (roi_weighted):

$$L = L_{global} + \lambda \cdot L_{ROI} \tag{6}$$

λ = 1.0 except in the filters = 16 sweep (0.5, 1.0, 2.0).croi_only (ablation):

$$L = L_{ROI} \tag{7}$$

with no supervision outside ROIs.

2.5. Training procedure and experimental matrix

The four filters = 32 runs used an RTX 4060; the five filters = 16 runs used Colab Pro. Table 1 lists the full experimental matrix; runs differed only in filters, loss, λ, and seed, which limits cross-capacity interpretation. torch/numpy/random were seeded, but strict CUDA determinism was not enabled. Checkpoints were saved every 5 epochs.

Table 1. Training configurations; λ applies only to roi_weighted

Run name	Filters	Loss	λ	Seed	Environment	Role	In H1–H3?
full_mae	32	mae	–	42	Local (RTX 4060)	Primary baseline	Yes
full_roi	32	roi_weighted	1.0	42	Local (RTX 4060)	Primary proposed	Yes
full_mae_seed7	32	mae	–	7	Local (RTX 4060)	H3 seed robustness	Yes
full_roi_seed7	32	roi_weighted	1.0	7	Local (RTX 4060)	H3 seed robustness	Yes
cap16_mae	16	mae	–	42	Colab Pro	H4 capacity baseline	No (val. only)
cap16_roi	16	roi_weighted	1.0	42	Colab Pro	H4 proposed; reused as λ=1.0 point	No (val. only)
cap16_roi_only	16	roi_only	–	42	Colab Pro	H5 ablation	No (val. only)

DOI: 10.54254/2753-8818/2027.MB37493

Table 1. (continued)

cap16_roi_lambda0p5	16	roi_weighted	0.5	42	Colab Pro	λ sensitivity	No (val. only)
cap16_roi_lambda2p0	16	roi_weighted	2.0	42	Colab Pro	λ sensitivity	No (val. only)

The official OpenKBP DoseEvaluator was used [4]. dose_score is the mean absolute voxel-wise dose error within the possible-dose region, reported in Gy. dvh_score pools valid patient–structure–metric errors for OAR D0.1cc/mean dose and PTV D99/D95/D1 using a NaN-aware mean. In contrast, the per-patient dvh_error used for hypothesis tests averages each patient's available metric errors first.

2.6. Statistical analysis plan

The main configuration (filters = 32, λ = 1.0) was not selected from the exploratory analyses. The plan was fixed before test-set evaluation but was not externally preregistered; H1 is therefore prespecified rather than formally confirmatory.

H1 (primary). dvh_error was averaged, for each patient, across seeds 42 and 7 within each condition. The paired difference was then calculated as proposed minus baseline. H1 needed both a two-sided Wilcoxon signed-rank test with $p < 0.05$, and a negative Hodges–Lehmann estimate [10, 11]. The means were reported descriptively. No multiplicity correction was needed for the single primary hypothesis.

H2 (descriptive). The same paired analyses were reported for dose_error. Without a prespecified non-inferiority margin, they do not establish preserved dose accuracy.

H3 (exploratory). Seeds 42 and 7 were reported separately; two seeds provide a robustness check, not a variance estimate.

H4 (exploratory). On the same 40 validation cases and seed 42, proposed-minus-baseline differences were compared directionally between filters = 32 and 16 for both errors. No interaction test was performed. Because capacity and training environment were confounded, this analysis assesses only agreement in effect direction, not a pure capacity effect or equality of effect sizes.

H5 (exploratory). At filters = 16, the loss ablation (mae/roi_weighted/roi_only) and λ sweep (0.5/1.0/2.0) were compared pairwise on validation data. Each three-condition family comprised six tests (three pairs × two metrics), adjusted by Holm's procedure [12]; raw and adjusted p values were reported.

Patient-level percentile bootstrap intervals used 10,000 resamples and seed 12345 [13]. Wilcoxon used zero_method="wilcox"; no outliers were excluded. Analysis halted if code-version identifiers mismatched.

2.7. Software and reproducibility safeguards

The network, loss, and training loop were implemented in PyTorch. Official OpenKBP data-loading, scoring, and basic I/O utilities were reused. This was a topology-matched reimplementation, not an exact reproduction: unlike the official baseline's all-voxel MAE, L_{global} was restricted to the possible-dose mask.

Each run stored a fixed CODE_VERSION in summary.json, and downstream analyses halted on a mismatch. SHA-256 fingerprints of training and analysis scripts were recorded when the plan was

DOI: 10.54254/2753-8818/2027.MB37493

frozen. These measures protect code consistency but do not replace external preregistration.

3. Results

Nine training configurations were studied in two U-Net capacities (32 and 16 initial convolutional filters) on the official OpenKBP data split (200 training / 40 validation / 100 test patients) [4]. The main analysis (Section 3.1) is the comparison of the baseline (voxel-wise MAE) and the proposed (MAE + per-structure ROI-weighted MAE, $\lambda = 1$) losses at the main model capacity (filters = 32) on the held-out 100-patient test set: for each patient the error is seed-averaged over two random initializations (seeds 42 and 7) and is analyzed with a two-sided Wilcoxon signed-rank test together with the Hodges–Lehmann estimator of the paired location shift. Sections 3.3–3.6 report the secondary and exploratory analyses (per-seed effects, capacity robustness, and an ablation of the loss components); they were performed on the 40-patient validation set and were not used to modify the primary model or its hyperparameters after the test-set evaluation.

3.1. Primary analysis: DVH accuracy (H1)

On the 100-patient test set, the proposed ROI-weighted loss reduced the DVH-metric error (dvh_error) relative to the MAE baseline for 99 of 100 patients (mean 3.58 Gy vs. 2.44 Gy), as shown in Figure 2a. The paired difference (proposed − baseline) had a Hodges–Lehmann (HL) estimate of −1.13 Gy (95% CI [−1.23, −1.04]) and a median of −1.15 Gy; the two-sided Wilcoxon signed-rank test rejected the null hypothesis of no shift ($p = 4.5 \times 10^{-18}$). Because both the significance test and the Hodges–Lehmann location estimate agreed in direction, H1 was supported under the prespecified decision rule (Section 2.6). A paired t-test on the same data (mean difference −1.14 Gy, $p = 1.3 \times 10^{-40}$) was consistent with this result but is reported only as a secondary sensitivity check, not as a co-primary test [14].

3.2. Descriptive analysis: dose accuracy (H2)

dvh_error is the error term most directly targeted by the proposed structure-weighted loss term; dose_error (the official challenge's whole-field dose accuracy score) was tracked descriptively to see whether the DVH improvement came at a cost to overall dose accuracy. On the same 100-patient, seed-averaged test-set comparison, and as shown in Figure 2b, dose_error also decreased under the proposed loss (mean 3.57 Gy vs. 3.36 Gy; 82/100 patients improved), with a Hodges–Lehmann estimate of −0.17 Gy (95% CI [−0.23, −0.13]) and Wilcoxon $p = 1.5 \times 10^{-11}$. No non-inferiority margin was prespecified for dose_error, so this result is reported as a descriptive effect size rather than a formal non-inferiority claim: within this dataset, the proposed loss did not trade dose accuracy for DVH accuracy — both improved together. These test-set comparisons are summarized in Table 2 and Figure 2.

Table 2. Primary (H1) and descriptive (H2) test-set results (n = 100)

Metric	Baseline mean (Gy)	Proposed mean (Gy)	Median diff (Gy)	HL diff [95% CI] (Gy)	n improved / 100	Wilcoxon p
dvh_error (H1, primary)	3.578	2.442	−1.145	−1.125 [−1.232, −1.035]	99	4.5×10^{-18}
dose_error (H2, descriptive)	3.565	3.362	−0.160	−0.175 [−0.225, −0.129]	82	1.5×10^{-11}

DOI: 10.54254/2753-8818/2027.MB37493

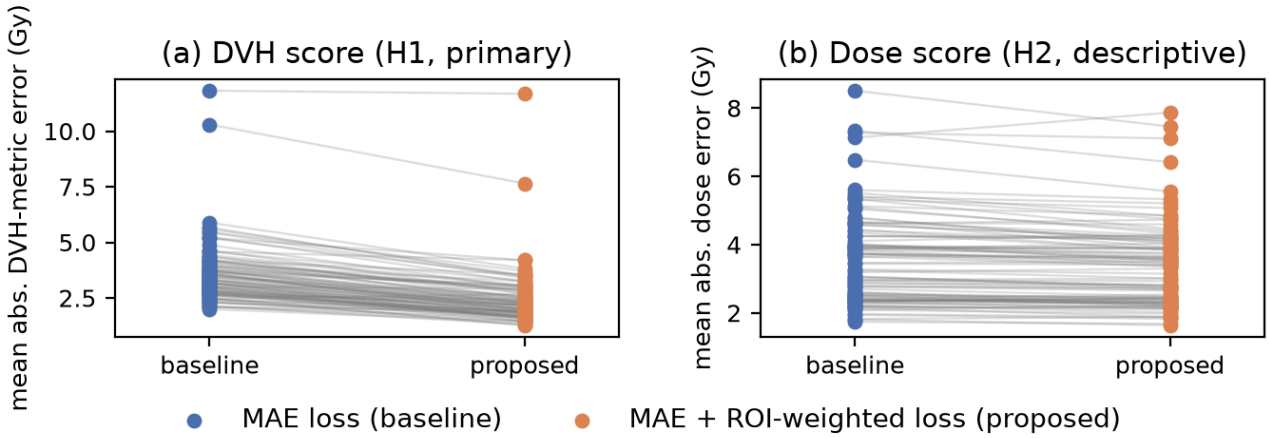

Figure 2. Paired per-patient comparison of dvh_error (H1, primary) and dose_error (H2, descriptive) between the baseline (MAE) and proposed (ROI-weighted) models on the official 100-patient test set, seed-averaged per patient (n = 100) (picture credit: original)

3.3. Illustrative case

Figures 3–4 illustrate the population-level effect reported in Sections 3.1–3.2. Patient pt_331, whose dvh_error improvement ranked 20th among the 100 test patients (i.e. a clearly visible but not the single most extreme effect), was selected as an illustrative case; population-level inference rests on the full n = 100 result in Section 3.1, not on this case. Figure 3 compares an axial dose-distribution slice through the OpenKBP reference dose, the baseline prediction, and the proposed prediction, shown respectively as Figures 3a, 3b, and 3c. Figure 4 shows the corresponding per-structure DVHs, with the three PTV dose levels appearing in Figures 4a–4c and the four OARs in Figures 4d–4g. The proposed model's DVH curves track the reference curves more closely than the baseline's through the steep dose-falloff ("shoulder") region of each PTV curve, as Figures 4a–4c show, and for the right parotid in Figure 4g, consistent with the aggregate dvh_error improvement reported above.

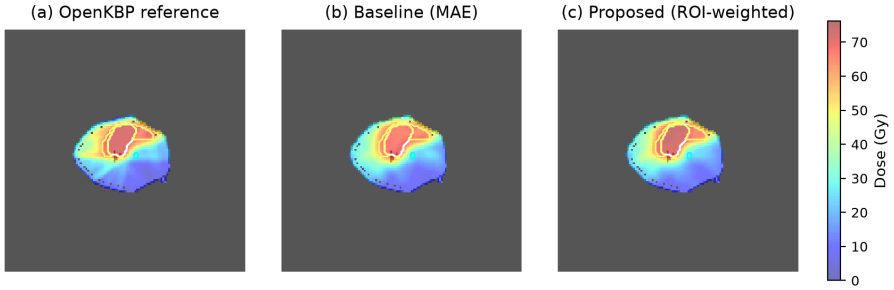

Figure 3. Illustrative case (pt_331, official test set): axial dose-distribution slice comparing the OpenKBP reference dose, the baseline (MAE) prediction, and the proposed (ROI-weighted) prediction. White contour: PTV70; yellow contour: PTV63; cyan contours: spinal cord and brainstem (picture credit: original)

DOI: 10.54254/2753-8818/2027.MB37493

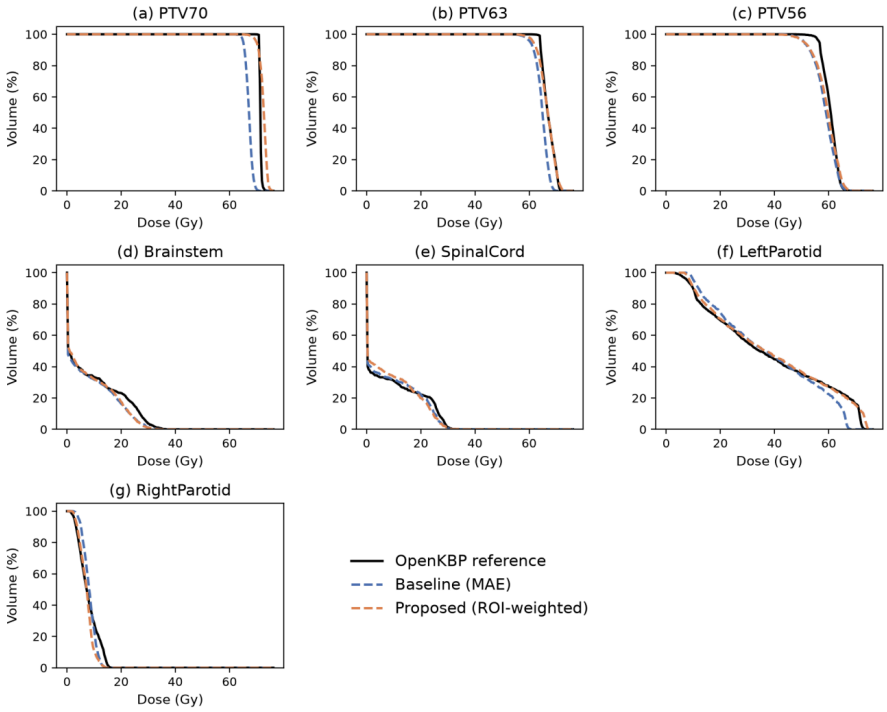

Figure 4. Illustrative case (pt_331, official test set): per-structure dose–volume histograms (PTV70, PTV63, PTV56, brainstem, spinal cord, left parotid, right parotid) comparing the reference dose (black, solid) against the baseline (blue, dashed) and proposed (orange, dashed) predictions (picture credit: original)

3.4. Seed robustness (H3)

The baseline-vs-proposed effect was reproduced independently under a second random initialization (seed 7), reported separately by seed rather than pooled into a false n = 200; Table 3 reports these per-seed results. Both seeds showed the same direction of effect for both metrics. On the validation set, notably, the baseline model's dvh_score varied considerably more across the two seeds (mean 3.35, SD 0.64 across seeds) than the proposed model's (mean 2.38, SD 0.07); the same pattern held for dose_score (baseline SD 0.17 vs. proposed SD 0.06), as Table 4 shows. With only two seeds this cannot be treated as a formal claim of reduced training variance, but it is a consistent and noteworthy secondary observation worth flagging for discussion.

Table 3. Per-seed paired results on the 100-patient test set

Seed	Metric	Baseline mean (Gy)	Proposed mean (Gy)	HL diff [95% CI] (Gy)	n improved / 100	Wilcoxon p
42	dvh_error	2.894	2.465	−0.406 [−0.508, −0.315]	80	1.6×10^{-11}
42	dose_error	3.401	3.314	−0.070 [−0.108, −0.034]	65	2.6×10^{-4}
7	dvh_error	4.262	2.418	−1.874 [−2.044, −1.695]	98	5.9×10^{-18}
7	dose_error	3.729	3.410	−0.295 [−0.379, −0.213]	82	1.9×10^{-11}

DOI: 10.54254/2753-8818/2027.MB37493

Table 4. Across-seed variability on the validation set, mean ± SD over the two seeds

Condition	dose_score mean ± SD	dvh_score mean ± SD	n seeds
Baseline (MAE)	3.511 ± 0.167	3.346 ± 0.643	2
Proposed (ROI-weighted)	3.312 ± 0.062	2.381 ± 0.067	2

3.5. Capacity robustness (H4)

To check whether the baseline-vs-proposed effect is specific to the capacity of the main model (filters = 32), a second capacity (filters = 16, single seed, trained on separate hardware) was compared on the 40-patient validation set; both capacities were evaluated on the same 40 official validation patients using the same scoring procedure. This analysis is exploratory: no test of $\Delta_{32} = \Delta_{16}$ is prespecified, given the different hardware, the single seed at filters = 16, and the absence of a formal interaction test; only the direction of each capacity's own effect is compared. As shown in Table 5, both capacities showed the same sign of effect (proposed better than baseline) for both dvh_error and dose_error, with an identical count of improved patients (27/40) for dvh_error at both capacities. This is consistent with the improvement not being an artifact specific to the capacity of the main model, but it does not establish a quantitative capacity-effect claim.

Table 5. Capacity comparison, 40-patient validation set

Capacity	Metric	Median Δ (Gy)	HL Δ [95% CI] (Gy)	n improved / 40	Direction vs. other capacity
filters=32	dvh_error	−0.122	−0.231 [−0.417, −0.058]	27	same
filters=32	dose_error	−0.028	−0.058 [−0.164, 0.012]	22	same
filters=16	dvh_error	−0.140	−0.138 [−0.255, −0.020]	27	same
filters=16	dose_error	−0.015	−0.020 [−0.085, 0.036]	21	same

3.6. Ablation of the global MAE term and λ sensitivity (H5)

To test whether the global MAE term contributes anything beyond the ROI-weighted term, an ablation condition (roi_only, i.e. loss = L_{ROI} with no L_{global} term) was trained at filters = 16 alongside the baseline and proposed conditions at the same capacity. As Figure 5a shows, removing the global term produced a large and highly significant increase in dose_error (mean dose_score 3.33 → 6.74 for baseline → roi_only, 0/40 patients improved, Wilcoxon $p = 1.8 \times 10^{-12}$, Holm-corrected $p = 1.1 \times 10^{-11}$), confirming that the global term is necessary to maintain overall dose accuracy: without it, the model is only supervised inside the ROI structures and its dose predictions degrade substantially outside them. dvh_score, shown in Figure 5b, was comparable between roi_only and the baseline (2.49 vs. 2.54, not significant after Holm correction) and somewhat worse than the full proposed loss (2.49 vs. 2.40). Tables 6 and 7 show numerically better DVH accuracy with the full proposed loss than with roi_only, although the difference was not statistically significant after Holm correction. Figure 5 shows the paired per-patient values underlying this comparison.

DOI: 10.54254/2753-8818/2027.MB37493

Table 6. Capacity/ablation trio, 40-patient validation set, filters=16

Condition	dose_score	dvh_score
MAE (baseline)	3.334	2.543
MAE + ROI-weighted (proposed, λ=1.0)	3.307	2.403
ROI-only (no global term)	6.736	2.490

Table 7. Pairwise tests for the ablation trio, Holm-Bonferroni corrected within this exploratory family of 6 comparisons

Comparison	Metric	Mean A	Mean B	n B better / 40	p (raw)	p (Holm)
MAE vs. ROI-weighted	dose_error	3.334	3.307	21	0.527	1.000
MAE vs. ROI-weighted	dvh_error	2.548	2.397	27	0.024	0.096
MAE vs. ROI-only	dose_error	3.334	6.736	0	1.8×10^{-12}	1.1×10^{-11}
MAE vs. ROI-only	dvh_error	2.548	2.501	28	0.048	0.144
ROI-weighted vs. ROI-only	dose_error	3.307	6.736	0	1.8×10^{-12}	1.1×10^{-11}
ROI-weighted vs. ROI-only	dvh_error	2.397	2.501	20	0.745	1.000

The ROI-term weight λ was varied over {0.5, 1.0, 2.0} at the same capacity (filters = 16) to check sensitivity to this hyperparameter. dvh_score decreased monotonically with λ over this range (2.49 [λ=0.5] → 2.40 [λ=1.0] → 2.35 [λ=2.0]); dose_score was essentially flat (3.36 [λ=0.5] → 3.31 [λ=1.0] → 3.36 [λ=2.0]); none of the pairwise comparisons reached significance after Holm correction, so this trend is reported descriptively and was not used to select λ for the main experiment — λ = 1.0 was fixed prospectively before this exploratory sweep was run. Table 8 and Figure 6 summarize these trends across λ; pairwise significance for each comparison is reported in Table 9.

Table 8. λ sensitivity, 40-patient validation set, filters=16

λ	dose_score	dvh_score
0.5	3.359	2.493
1.0	3.307	2.403
2.0	3.359	2.351

Table 9. Pairwise tests for the λ sweep, Holm-Bonferroni corrected within this exploratory family of 6 comparisons

Comparison	Metric	p (raw)	p (Holm)
λ=0.5 vs. λ=1.0	dose_error	0.183	0.732
λ=0.5 vs. λ=1.0	dvh_error	0.390	0.780
λ=0.5 vs. λ=2.0	dose_error	0.775	0.780
λ=0.5 vs. λ=2.0	dvh_error	0.060	0.301
λ=1.0 vs. λ=2.0	dose_error	0.026	0.155

DOI: 10.54254/2753-8818/2027.MB37493

Table 9. (continued)

λ=1.0 vs. λ=2.0	dvh_error	0.259	0.777

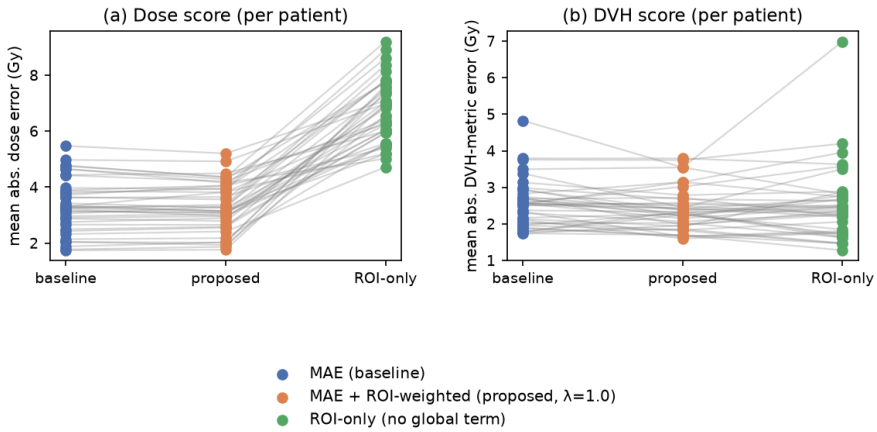

Figure 5. Capacity/ablation trio at filters=16 (validation set): paired per-patient dose_error and dvh_error for MAE (baseline), MAE + ROI-weighted (proposed), and ROI-only (ablation) (picture credit: original)

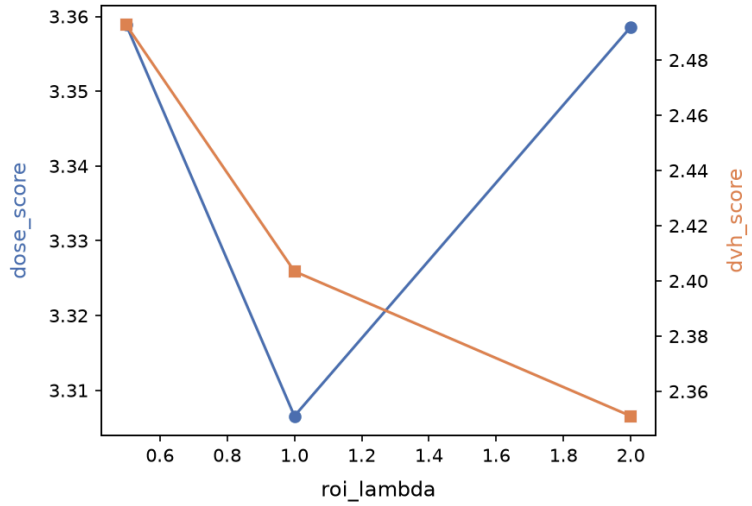

Figure 6. λ sensitivity at filters=16 (validation set): dose_score and dvh_score as a function of the ROI-term weight $\lambda \in \{0.5, 1.0, 2.0\}$ (picture credit: original)

4. Discussion

The proposed ROI-weighted loss reduced DVH-metric error on the held-out test set (Hodges–Lehmann paired difference [proposed − baseline]: −1.13 Gy, 95% CI [−1.23, −1.04]; mean dvh_error 3.58 → 2.44 Gy, 99/100 patients improved, Wilcoxon $p = 4.5\times10^{-18}$); H1 was supported under the prespecified decision rule. Critically, this was not at the expense of overall dose accuracy: dose_error decreased as well (H2, descriptive; mean 3.57 → 3.36 Gy, HL −0.17 Gy). A structure-level reweighting of the loss therefore improved the metric it directly targets without degrading the metric it does not. Knowledge-based planning methods commonly predict DVHs, dose–volume

DOI: 10.54254/2753-8818/2027.MB37493

metrics, or voxel-wise dose distributions to guide treatment planning [1]. Accordingly, an improvement in DVH-focused error is more directly aligned with a principal class of KBP outputs than whole-field average error alone.

Why? The ablation (H5) clarifies why. Dropping the global MAE term entirely (roi_only) left dose_score substantially worse (3.33 → 6.74 Gy, 0/40 patients improved) while leaving dvh_score roughly unchanged from baseline — the model loses all supervision outside the ten labeled ROIs and its predictions degrade across the rest of the field. The proposed loss keeps the global term and adds the ROI term on top, and it is this combination, not the ROI term alone, that improved both metrics together. The mechanism is best read as redistributing gradient emphasis toward clinically important structures without removing the field-wide supervisory signal, rather than as a substitute for it. This interpretation is consistent with prior structure-informed objectives based on per-structure moments [5], composite structure-aware DVH and spatial-gradient terms [6], or fixed ROI-weighted MSE [7]. These differ from the present per-patient-normalized, possible-dose-mask-intersected ROI MAE.

The effect was not fragile to the nuisance factors that could be checked. H3 reproduced the same direction of effect under an independent seed, and — a secondary but notable observation, not treated as a formal claim with only two seeds — the across-seed variability of the proposed model is smaller than that of the baseline on the validation set (dvh_score SD 0.07 vs. 0.64; dose_score SD 0.06 vs. 0.17), so the ROI term may not only shift its average outcome but also stabilize training. H4 found the same direction of effect at a smaller capacity (filters = 16) trained on separate hardware, and the improved-patient count (27/40) for dvh_error is identical at both capacities; this is evidence against the improvement being an artifact of the main model's specific capacity, though the different hardware and single seed at filters = 16 keep this a directional check, not a quantitative capacity-effect estimate. The λ sweep shows a monotonic but non-significant trend toward larger dvh_score gains at higher λ (2.49 → 2.40 → 2.35 for λ = 0.5/1.0/2.0) with dose_score essentially flat; λ = 1.0 was fixed before this sweep was run and is a reasonable default, but the sweep does not establish it as optimal.

There are several limitations that bound these conclusions. Firstly, the reference dose is the OpenKBP dataset's own synthetic automated-planning-pipeline output rather than an original clinical dose (Section 2.1); an improvement against this reference is evidence of agreement with that planning pipeline's output, rather than a directly verified clinical-quality improvement. dose_score and dvh_score are the official challenge's surrogate accuracy metrics and not a full deliverability or plan-quality review; and no predicted dose distribution here was assessed by a clinician. Moreover, all the results come from one dataset, one anatomical site (head-and-neck), and one architecture family, so generalization to other sites, planning pipelines, or network designs is not tested; a recent systematic review of head-and-neck dose-prediction models reported this same narrow-dataset pattern, together with a general lack of external validation, across the published literature [15]. The filters = 16 arm — the underlying H4 and H5 — used a single seed on different hardware from the two-seed, single-hardware main experiment, and the λ sweep covered only three values at that one capacity; a wider, main-capacity sweep would more firmly establish the effect of λ. Finally, the dose_error result of H2 is descriptive: no non-inferiority margin was prespecified, so "did not trade off" is a characterization of the observed effect size rather than a formally tested claim. Nor was plan deliverability or complexity assessed — a dose distribution can score well on dose_score/dvh_score while still being difficult to deliver with a real linear accelerator, and this study does not speak to that question.

DOI: 10.54254/2753-8818/2027.MB37493

These limitations point directly to future work: evaluating the λ sweep at the main model capacity, testing the approach on additional cohorts and treatment sites, and comparing the per-patient-normalized ROI term used here against the moment-based and fixed-schedule structure weightings used in prior work under a common evaluation protocol. Within the scope tested, though, a simple, easily implemented modification to the training loss — no architecture change, no extra inference cost — produced a large, statistically robust, direction-consistent improvement in the metric it targeted, with no observed cost to the metric it did not.

5. Conclusion

This work reimplemented a 3-D U-Net dose predictor on the official OpenKBP benchmark and tested, under a single prespecified primary hypothesis, whether adding a per-structure ROI-weighted term to a standard masked-MAE loss improves DVH accuracy. On the held-out 100-patient test set, the proposed loss reduced dvh_error, with a Hodges–Lehmann paired difference (proposed − baseline) of −1.13 Gy (95% CI [−1.23, −1.04]; $p = 4.5\times10^{-18}$), relative to the MAE baseline, with dose_error also improving rather than being traded off. The effect held in direction across an independent seed and a smaller model capacity, and an ablation showed the global MAE term remains necessary — the ROI term supplements it rather than replacing it. Removing the global term entirely, as the ablation showed, left whole-field dose accuracy far worse even though the labeled structures themselves were largely unaffected, which is why the two terms need to stay together rather than one substituting for the other. Because the change lives entirely in the loss function, it needs no new architecture, no extra parameters, and no added inference cost, so it should be straightforward to try on other dose-prediction pipelines beyond the one reimplemented here. The reference dose's synthetic origin, the single dataset and anatomical site, and the absence of clinician review of the predicted plans bound how far these results generalize. The next steps follow directly from where this evaluation stopped short: testing the approach on other cohorts and treatment sites, having a clinician assess the resulting plans rather than relying on the challenge's surrogate metrics alone, and running the λ sweep at the main model capacity instead of only the reduced one used here. Within that scope, a small, architecture-agnostic change to the training loss produced a statistically robust, directionally consistent improvement in dose-prediction accuracy.

References

[1] Momin, S., Fu, Y., Lei, Y., Roper, J., Bradley, J. D., Curran, W. J., Liu, T., & Yang, X. (2021). Knowledge-based radiation treatment planning: A data-driven method survey. *Journal of Applied Clinical Medical Physics, 22*(8), 16–44. https://doi.org/10.1002/acm2.13337

[2] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015 (Lecture Notes in Computer Science, Vol. 9351)* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28

[3] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016). 3D U-Net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016 (Lecture Notes in Computer Science, Vol. 9901)* (pp. 424–432). Springer. https://doi.org/10.1007/978-3-319-46723-8_49

[4] Babier, A., Zhang, B., Mahmood, R., Moore, K. L., Purdie, T. G., McNiven, A. L., & Chan, T. C. Y. (2021). OpenKBP: The open-access knowledge-based planning grand challenge and dataset. *Medical Physics, 48*(9), 5549–5561. https://doi.org/10.1002/mp.14845

[5] Jhanwar, G., Dahiya, N., Ghahremani, P., Zarepisheh, M., & Nadeem, S. (2022). Domain knowledge driven 3D dose prediction using moment-based loss function. *Physics in Medicine & Biology, 67*(18), Article 185017. https://doi.org/10.1088/1361-6560/ac8d45

DOI: 10.54254/2753-8818/2027.MB37493

[6] Mak, P. C. Y., Kong, L., & Chan, L. W. C. (2026). Deep-learning-based 3D dose distribution prediction for VMAT lung cancer treatment using an enhanced UNet3D architecture with composite loss functions. *Bioengineering, 13*(5), Article 490. https://doi.org/10.3390/bioengineering13050490

[7] Ogawa, S., Kadoya, N., Kato, T., Kato, R., Tozuka, R., Narita, Y., Oyama, S., Murakami, M., & Jingu, K. (2026). Deep learning-based dose prediction in proton beam therapy for hepatocellular carcinoma: Comparison of network architectures and loss functions. *Journal of Radiation Research, 67*(3), 369–383. https://doi.org/10.1093/jrr/rrag026

[8] Maniscalco, A., Liang, X., Lin, M.-H., Jiang, S., & Nguyen, D. (2023). Single patient learning for adaptive radiotherapy dose prediction. *Medical Physics, 50*(12), 7324–7337. https://doi.org/10.1002/mp.16799

[9] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015)*. https://arxiv.org/abs/1412.6980

[10] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin, 1*(6), 80–83. https://doi.org/10.2307/3001968

[11] Hodges, J. L., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics, 34*(2), 598–611. https://doi.org/10.1214/aoms/1177704172

[12] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70. https://www.jstor.org/stable/4615733

[13] Rousselet, G. A., Pernet, C. R., & Wilcox, R. R. (2021). The percentile bootstrap: A primer with step-by-step instructions in R. *Advances in Methods and Practices in Psychological Science, 4*(1). https://doi.org/10.1177/2515245920911881

[14] Student. (1908). The probable error of a mean. *Biometrika, 6*(1), 1–25. https://doi.org/10.1093/biomet/6.1.1

[15] Zamanian, M., Kavehpoor, M. A., Soltaninejad, A. M., Abedi, I., & Tavakkoli, M. B. (2025). Deep learning for head and neck radiation dose prediction: A systematic review and meta-analysis. *Scientific Reports, 15*, Article 45766. https://doi.org/10.1038/s41598-025-28442-3