Closing the loop: fully automated radiotherapy planning for head and neck cancer

Gao, R.

4

Clinical DVH Metrics as a Loss Function for 3D Dose Prediction in Head and Neck Radiotherapy

This chapter was adapted from:

Ruochen Gao, Marius Staring, and Frank Dankers, "Clinical DVH metrics as a loss function for 3D dose prediction in head and neck radiotherapy", accepted for publication in the International Journal of Computer Assisted Radiology and Surgery (IJCARS), CARS 2026 Special Issue.

Abstract

Purpose Deep-learning-based three-dimensional (3D) dose prediction has become an important component of automated radiotherapy workflows. However, most existing models are trained using voxel-wise regression losses, which are poorly aligned with clinical plan evaluation criteria that rely on dose-volume histogram (DVH)-derived metrics. This study aims to develop a clinically guided loss formulation that directly optimizes clinically used DVH metrics while remaining computationally efficient for head and neck (H&N) dose prediction.

Methods We propose a clinical DVH metric loss (CDM loss) that jointly incorporates differentiable formulations of common *D–metrics* and differentiable surrogates of *V–metrics*. The loss is driven by a JSON-based clinical plan evaluation template to ensure alignment with clinical planning criteria. In addition, we introduce a lossless bit-mask encoding scheme to efficiently represent a large number of overlapping ROIs, substantially improving training efficiency. The method was evaluated on a retrospective cohort of 174 H&N patients treated with VMAT, using a temporal split for training ($n = 137$) and testing ($n = 37$).

Results Compared with MAE- and DVH-based losses, the proposed CDM loss substantially improved clinical target coverage, and satisfied all predefined clinical constraints. Using a standard 3D U-Net, the PTV Score was reduced from 1.544 (MAE) to 0.491 (MAE+CDM), while consistently satisfying all PTV clinical constraints. OAR sparing remained comparable or improved. Bit-mask ROI encoding reduced average training epoch time from 241 s to 43 s (82.2% reduction) and lowered peak GPU memory usage. Notably, with the CDM loss, a standard 3D U-Net achieved performance comparable to or better than more complex state-of-the-art architectures.

Conclusion Directly optimizing clinically used DVH metrics enables 3D dose predictions that are better aligned with clinical treatment planning criteria than conventional only voxel-wise or DVH-curve supervision. The proposed CDM loss, combined with efficient ROI bit-mask encoding, provides a practical and scalable framework for H&N dose prediction.

4.1 Introduction

Radiotherapy (RT) is widely used in cancer treatment to deliver sufficient radiation to eradicate tumor cells while protecting surrounding healthy tissues. Conventional treatment planning directly optimizes beam parameters within the treatment planning system (TPS), often requiring multiple rounds of manual refinements and substantial planner expertise. This process is highly time-consuming, especially for anatomically complex sites. To improve efficiency and consistency, modern workflows increasingly introduce dose prediction as an intermediate step, providing an anatomy-informed estimate of a clinically realistic three-dimensional (3D) dose distribution. This predicted dose can then be used to guide subsequent TPS optimization through a dose mimicking process, in which a deliverable treatment plan is optimized to closely reproduce the predicted 3D dose distribution [85].

Among all treatment sites, head and neck (H&N) radiotherapy is particularly challenging due to its intricate anatomy, a large number of critical organs-at-risk (OARs), and the need for steep dose gradients between planning target volumes (PTVs) and nearby healthy organs [9]. These complexities make high-quality optimization particularly demanding and contribute to considerable variability in plan quality across clinicians and institutions [21]. Therefore, developing accurate and automated 3D dose prediction models tailored for H&N treatment is essential to reduce planning workload and promote standardization in clinical practice.

Early dose prediction approaches relied on atlas-based and knowledge-based planning, which generated 3D dose estimates from prior clinical plans and handcrafted anatomical features [103]. Although these methods improve planning consistency, they are limited in capturing the complex nonlinear relationship between anatomy and dose. Recently, deep learning (DL) methods have shown strong potential for learning complex mappings between patient anatomy and clinical dose patterns. Most DL-based methods use 3D convolutional or transformer-based U-shaped architectures [91, 93, 24, 28], taking CT and region–of–interest (ROI) masks as input to generate 3D voxel-wise dose predictions, as shown in Fig. 4.1. These approaches generally achieve high voxel-wise accuracy because they are trained with regression losses such as mean absolute error (MAE) or mean squared error (MSE), which encourage global similarity to clinical dose distributions. However, voxel-wise accuracy is not the primary criterion used in clinical plan evaluation. Instead, clinicians focus on achieving adequate planning target volume (PTV) coverage and sufficient organ-at-risk (OAR) sparing, which are typically assessed using dose–volume histogram (DVH)-derived metrics. Because standard regression losses do not directly optimize these clinical objectives, DL-based models may achieve high numerical accuracy, yet still perform sub-optimally from a clinical standpoint.

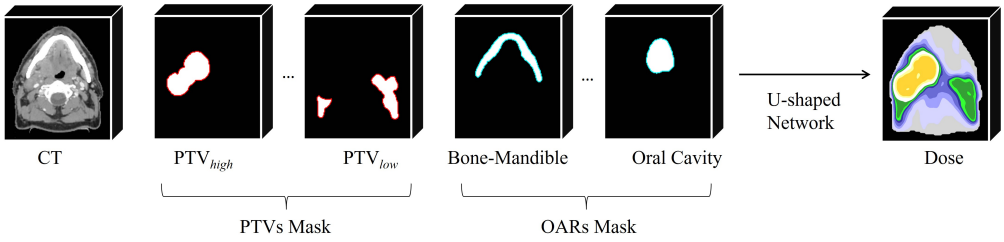


Figure 4.1: Workflow of a typical deep-learning-based H&N dose prediction method. The planning CT and masks of the PTVs and OARs are provided as inputs to a U-shaped neural network to generate a 3D voxel-wise dose distribution. In H&N cases, the PTVs are typically assigned two or three dose levels, while the number of OARs is relatively large compared to other treatment sites.

Table 4.1: Definitions of common *D*– and *V*–*metrics* in H&N RT.

Parameter	**Definition**
$D_{x\%}$	Minimum dose received by x% of the ROI volume.
$D_{x\ \text{cc}}$	Minimum dose received by the hottest x cc of the ROI.
D_{max}	Maximum dose within the ROI.
D_{min}	Minimum dose within the ROI.
D_{mean}	Mean dose within the ROI.
$V_{x\%}$	Volume percentage of an ROI receiving at least x% of the prescription dose.
$V_{x\ \text{Gy}}$	Volume percentage of an ROI receiving at least x Gy.

To bridge this gap, several studies have incorporated differentiable DVH losses, which allow comparison of predicted and ground-truth DVH curves during training [95, 104, 26, 27]. In clinical practice, however, plan evaluation is typically based on a collection of dose metrics rather than full DVH curves. For H&N RT, these metrics commonly include both *D*–*metrics* (e.g., $D_{x\%}$, the minimum dose received by x% of an ROI) and *V*–*metrics* (e.g., $V_{x\%}$, the volume of an ROI receiving at least x% of the prescription dose), as summarized in Table 4.1. While *D*–*metrics* are differentiable and can be incorporated into training [26, 27], *V*–*metrics* contain threshold operations that are inherently non-differentiable, posing challenges for the gradient-based training framework. Therefore, effective methods for directly optimizing *V*–*metrics* remain limited, and a comprehensive framework that jointly supports both *D*– and *V*–*metrics* is still needed.

Another practical challenge arises from the large number of OARs in H&N anatomy. Unlike DL-based medical imaging segmentation models, which often train on small 3D patches, dose–prediction networks typically require the full 3D volume as input to guarantee spatial dose continuity. Current methods generally implement these ROIs as one input channel per ROI. As a result, when many OARs are included, the

number of input channels grows rapidly, leading to substantial preprocessing overhead, higher CPU—to—GPU data transfer cost, increased GPU memory usage, and longer training time. Although merging all ROIs into a single indexed mask could reduce input dimensionality, this approach faces two major issues: (1) many OARs overlap spatially (e.g., the brain and brainstem overlap), making direct merging impossible; and (2) the numerical labels introduce unintended ordinal bias (e.g., a label "9" may be treated differently from a label "1" purely because it is larger), even though ROI labels should carry no numeric significance. This unintended numerical meaning can mislead the network during training and negatively affect learning [105]. As a result, prior studies often limit the input to a subset of OARs [22, 93, 24, 28, 26, 27], which is restrictive because clinical evaluation typically involves a much broader set of OARs. This challenge highlights the need for models that can efficiently incorporate all clinically relevant ROIs while maintaining computational efficiency.

To address these limitations, we propose a clinical DVH metric loss (CDM loss), together with an efficient bit–mask ROI encoding strategy for 3D H&N dose prediction. Our framework enables direct optimization of the full set of clinically defined dose evaluation metrics while maintaining computational efficiency. Our main contributions are summarized as follows:

- We introduce a clinical DVH metric loss (CDM loss) that combines differentiable formulations for *D–metrics* with differentiable approximations for *V–metrics*. By directly optimizing the metrics used in clinical plan evaluation, we show that even a standard 3D U-Net can achieve clinically satisfactory dose predictions.
- We develop an efficient lossless bit–mask encoding for ROI representation that reduces CPU preprocessing, CPU—to—GPU data transfer, and GPU memory usage through lightweight, on-demand decoding. This strategy substantially accelerates training, while supporting the full set of OARs required for clinical evaluation.
- We organize the clinical evaluation dose metrics used by CDM loss function through a simple JSON-based template, as summarized in Table 4.2. This template specifies the ROIs and their associated dose metrics in a clear, structured, and easily updated format, ensuring that the training objectives remain aligned with clinical evaluation criteria used in a given institute.

4.2 Materials

4.2.1 Dataset

In this study, data were collected retrospectively from 174 H&N cancer patients treated at Leiden University Medical Center from February 2021 to July 2025. The study

Table 4.2: The H&N RT clinical plan evaluation template used in our institution. The PTVs are denoted by their prescribed dose levels (e.g., $PTV_{54.25}$ for 54.25 Gy and PTV_{70} for 70 Gy). ROIs with "L/R" denote paired organs and are evaluated independently for the left and right sides. Each ROI is evaluated using one or more dose metrics. "Aim" represents the desired planning goal, while "Constraint" denotes a mandatory criterion. The template is implemented via a JSON file.

ROI	**Metric**	**Aim**	**Constraint**
$PTV_{54.25}$	$V_{95\%}$		$\geq 98\%$
$PTV_{54.25}$	D_{mean}	$\leq 102\%$	
PTV_{70}	$V_{95\%}$		$\geq 98\%$
PTV_{70}	$D_{0.03\text{cc}}$	$\leq 107\%$	$\leq 110\%$
PTV_{70}	D_{mean}	$\leq 102\%$	
SpinalCord	$D_{0.03\text{cc}}$		$\leq 50\,\text{Gy}$
SpinalCord + 3 mm	$D_{0.03\text{cc}}$		$\leq 52\,\text{Gy}$
Brainstem_Surf	$D_{0.03\text{cc}}$		$\leq 60\,\text{Gy}$
Brainstem_Core	$D_{0.03\text{cc}}$		$\leq 54\,\text{Gy}$
Brain	$D_{0.03\text{cc}}$	$\leq 65\,\text{Gy}$	
Brain	$D_{2\%}$	$\leq 70\,\text{Gy}$	
Cochlea_(L/R)	D_{mean}	$\leq 45\,\text{Gy}$	
Parotid_(L/R)	D_{mean}	$\leq 28\,\text{Gy}$	
Glnd_Submand_(L/R)	D_{mean}	$\leq 35\,\text{Gy}$	
Oral_Cavity	D_{mean}	$\leq 28\,\text{Gy}$	
Musc_Constrict_S	D_{mean}	$\leq 40\,\text{Gy}$	
Musc_Constrict_M	D_{mean}	$\leq 40\,\text{Gy}$	
Musc_Constrict_I	D_{mean}	$\leq 40\,\text{Gy}$	
Cricopharyngeus	D_{mean}	$\leq 40\,\text{Gy}$	
Larynx_SG	D_{mean}	$\leq 40\,\text{Gy}$	
Glottic_Area	D_{mean}	$\leq 40\,\text{Gy}$	
Bone_Mandible	$D_{2\%}$	$\leq 70\,\text{Gy}$	
Bone_Mandible-PTV	$D_{2\%}$	$\leq 50\,\text{Gy}$	
Eye_(L/R)	$D_{0.03\text{cc}}$	$\leq 35\,\text{Gy}$	
Lens_(L/R)	$D_{0.03\text{cc}}$	$\leq 6\,\text{Gy}$	
Pituitary	D_{mean}	$\leq 20\,\text{Gy}$	
OpticNrv_(L/R)	$D_{0.03\text{cc}}$	$\leq 55\,\text{Gy}$	

was approved by the Medical Ethics Committee of Leiden, The Hague, and Delft (approval number G21.142, October 15, 2021). The cohort covered a range of tumor sites, including the oropharynx, hypopharynx, nasopharynx, larynx, and oral cavity. All patients were treated on Elekta Synergy linear accelerators (Elekta, Stockholm, Sweden) using 6 MV dual-arc, full-rotation volumetric modulated arc therapy (VMAT) and all RT plans were generated in RayStation$^{\text{TM}}$ (RaySearch Laboratories, Stockholm,

Sweden).

To better reflect real-world clinical application, in which a model trained on historical data is used to make predictions for future patients, the dataset was divided chronologically. Patients treated between February 2021 and July 2024 ($n = 137$) were assigned to the training cohort, while those treated between August 2024 and July 2025 ($n = 37$) formed the test cohort, enabling temporal (out-of-time) validation on later cases.

The original CT images had an average resolution of $1.2 \times 1.2 \times 2\,\mathrm{mm}^3$, whereas the dose grid resolution was fixed at $2 \times 2 \times 2\,\mathrm{mm}^3$. To ensure uniform spatial resolution, all CT scans were resampled to $2 \times 2 \times 2\,\mathrm{mm}^3$. Cropping or padding were performed around the treatment isocenter, defined as the geometric center of the planned radiotherapy beams, to obtain a uniform volume size of $(160, 192, 256)$. The same preprocessing pipeline was applied to all ROIs.

For intensity normalization, CT values were clipped to the range $[-1024, 1024]$ and then normalized to $[0, 1]$ using min–max normalization. Dose values were normalized by dividing by 70 (Gy), the highest prescription dose in our cohort.

4.2.2 Evaluation metrics

To quantitatively evaluate model performance, three complementary evaluation metrics are defined: PTV Score, OAR Score, and Dose Score. The PTV Score and OAR Score are directly derived from the clinical plan evaluation template used in our institution (Table 4.2), which specifies the clinical dose evaluation metrics for each ROI. These two scores are reported separately because they represent different clinical objectives and optimization priorities.

Let $\mathcal{P}$ denote the set of all PTVs, and $\mathcal{O}$ denote the set of all OARs. For each ROI $r \in \mathcal{P} \cup \mathcal{O}$, a corresponding set of clinical evaluation dose metrics $\mathcal{M}_r$ is defined according to Table 4.2. For example, $\mathcal{M}_r = \{V_{95\%},\ D_{\text{mean}},\ D_{0.03\text{cc}}\}$ for PTV_{70}, and $\mathcal{M}_r = \{D_{0.03\text{cc}},\ D_{2\%}\}$ for Brain. For each metric $m \in \mathcal{M}_r$, let $M_{r,m}^{\text{pred}}$ and $M_{r,m}^{\text{gt}}$ denote the predicted and ground truth (clinical) values, respectively.

The PTV Score is defined as:

$$\text{PTV Score} = \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \frac{1}{|\mathcal{M}_p|} \sum_{m \in \mathcal{M}_p} \left| M_{p,m}^{\text{pred}} - M_{p,m}^{\text{gt}} \right|, \tag{4.1}$$

where all PTV metrics are normalized to the prescribed dose and expressed in percentage (%). Similarly, the OAR Score is defined as:

$$\text{OAR Score} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \frac{1}{|\mathcal{M}_o|} \sum_{m \in \mathcal{M}_o} \left| M_{o,m}^{\text{pred}} - M_{o,m}^{\text{gt}} \right|, \tag{4.2}$$

where all OAR metrics are measured in absolute dose (Gy).

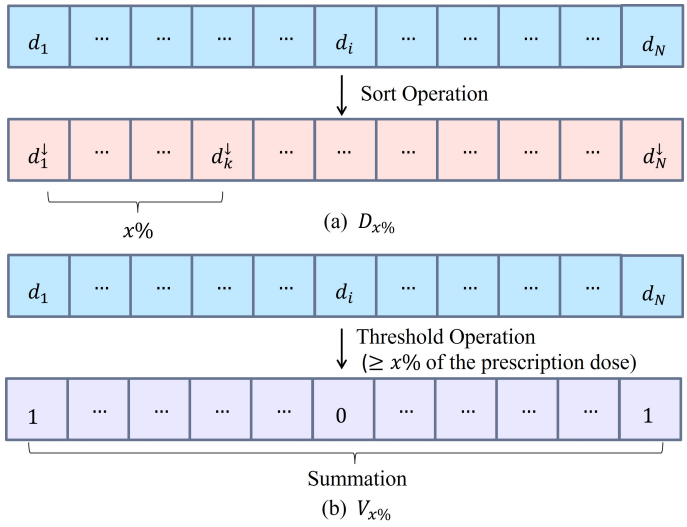


Figure 4.2: (a) The 3D dose distribution within an ROI is flattened into a one-dimensional dose array $\{d_i\}_{i=1}^{N}$. After a descending sorting operation, we get $d_1^{\downarrow} \geq d_2^{\downarrow} \geq \cdots \geq d_N^{\downarrow}$. A *D–metric*, such as $D_{x\%}$, is obtained by retrieving the dose value $d_k^{\downarrow}$ at the index k that corresponds to the top x% portion of the ROI volume in the sorted dose array. (b) Using the same flattened dose array $\{d_i\}_{i=1}^{N}$, a *V–metric* such as $V_{x\%}$ is computed by applying a dose threshold equal to x% of the prescription dose, summing the number of voxels that exceed this threshold, and computing the corresponding volume percentage. For example, $V_{95\%}$ of PTV_{70} denotes the fraction of the PTV_{70} volume receiving at least 95% of the 70 Gy dose level.

The Dose Score is also reported to quantify the overall voxel-wise difference between the predicted and ground truth (clinical) 3D dose distributions. Let Ω denote the set of all voxels within the dose grid, and $D^{\text{pred}}(v)$ and $D^{\text{gt}}(v)$ be the predicted and clinical dose at voxel $v \in \Omega$. The Dose Score is defined as:

$$\text{Dose Score} = \frac{1}{|\Omega|} \sum_{v \in \Omega} \left| D^{\text{pred}}(v) - D^{\text{gt}}(v) \right|, \tag{4.3}$$

reported in Gy. Although this metric has limited direct clinical relevance, it is widely adopted in the literature and facilitates consistent comparison with prior studies.

4.3 Methods

4.3.1 CDM loss formulation

Based on the clinical plan evaluation criteria summarized in Table 4.2, we formulate the proposed CDM loss that directly optimizes the specified dose metrics. It incorporates both *D–* and *V–metrics*, allowing end-to-end training guided by clinical evaluation

criteria. The following sections describe the differentiable formulation of *D–metrics* and the surrogate-based differentiable approximation of *V–metrics*.

4.3.1.1 Differentiable formulation of *D–metrics*

As shown in Fig. 4.2(a), given a set of voxel doses $\{d_i\}_{i=1}^{N}$ within an ROI of size N, sorted in descending order as $d_1^{\downarrow} \geq d_2^{\downarrow} \geq \cdots \geq d_N^{\downarrow}$, the minimum dose received by the hottest x% of the ROI volume, is denoted as:

$$D_{x\%} = d_k^{\downarrow}, \text{ where } k = \left\lceil \frac{x}{100} N \right\rceil. \tag{4.4}$$

Thus, $D_{x\%}$ corresponds to the k-th largest dose value, which can be found at index k after sorting. Similarly, the minimum dose received by the hottest x cc of the ROI, denoted as $D_{x\text{ cc}}$, is computed in an analogous manner:

$$D_{x\text{ cc}} = d_k^{\downarrow}, \text{ where } k = \left\lceil \frac{x}{V_{\text{voxel}}} \right\rceil, \tag{4.5}$$

where V_{voxel} denotes the volume of a single voxel (in cc). Both $D_{x\%}$ and $D_{x\text{ cc}}$ can therefore be interpreted as quantile-based metrics, differing only in whether the cutoff is defined by a fractional or absolute volume.

In addition to these quantile metrics, maximum and minimum dose within an ROI can be expressed naturally in the same framework:

$$D_{\max} = d_1^{\downarrow}, \quad D_{\min} = d_N^{\downarrow}. \tag{4.6}$$

Although sorting indices are non-differentiable, deep learning frameworks such as `PyTorch`[1] allow gradients to propagate through the sorted values via gather/scatter operations. Following prior work [26, 27], we use a top-k algorithm that computes the required quantiles without explicitly sorting all voxels, offering both differentiability and computational efficiency.

Finally, the mean dose $D_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} d_i$ is inherently differentiable.

4.3.1.2 Differentiable surrogate of *V–metrics*

V–metrics, including $V_{x\%}$ and $V_{x\text{ Gy}}$, describe the fraction of an ROI receiving at least a specified dose level. As illustrated in Fig. 4.2(b), $V_{x\%}$ denotes the fraction of the ROI receiving at least x% of the prescription dose. Given the prescription dose D_{presc}, the corresponding threshold is $T = xD_{\text{presc}}/100$ Gy. The computation of $V_{x\%}$ is equivalent to counting the number of voxels with dose $d_i \geq T$, normalized by the total number of voxels N:

$$V_{x\%} = \frac{1}{N}\sum_{i=1}^{N} H(d_i - T), \tag{4.7}$$

[1] `https://pytorch.org/`

where $H(\cdot)$ is the non-differentiable Heaviside step function. The same principle applies to $V_{x\ \mathrm{Gy}}$, differing only in that the threshold is specified in absolute dose.

To enable gradient-based training, we approximate H with a logistic sigmoid $\sigma_\alpha(t) = 1/(1+e^{-\alpha t})$, with slope parameter $\alpha > 0$:

$$V_{x\%}^{\text{approx}} = \frac{1}{N}\sum_{i=1}^{N} \sigma_\alpha(d_i - T). \tag{4.8}$$

A larger α causes σ_α to more closely approximate the hard threshold, but could also increase the risk of gradient saturation: when α is too large, most voxels have $|\alpha(d_i - T)| \gg 1$, placing them in the flat tails of the sigmoid where $\sigma'_\alpha(t) \approx 0$. Consequently, only voxels very close to T contribute meaningful gradients, which may destabilize training or slow convergence. Therefore, rather than arbitrarily increasing α, we seek the smallest value that ensures sufficient approximation accuracy while mitigating the risk of gradient saturation.

4.3.1.3 Selection of α via a bound on the approximation error

Let $z_i = d_i - T$. We first quantify how closely the smooth surrogate $\sigma_\alpha(z)$ matches the hard indicator $H(z)$ at the voxel level. The mean pointwise approximation error is defined as:

$$\delta(\alpha) = \frac{1}{N}\sum_{i=1}^{N} |H(z_i) - \sigma_\alpha(z_i)|. \tag{4.9}$$

Since

$$V_{x\%} = \frac{1}{N}\sum_{i=1}^{N} H(z_i), \qquad V_{x\%}^{\text{approx}}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} \sigma_\alpha(z_i), \tag{4.10}$$

the triangle inequality implies that the $V_{x\%}$ approximation error is always bounded by

$$\left|V_{x\%} - V_{x\%}^{\text{approx}}(\alpha)\right| = \left|\frac{1}{N}\sum_{i=1}^{N} (H(z_i) - \sigma_\alpha(z_i))\right| \le \delta(\alpha). \tag{4.11}$$

We now derive an upper bound for $\delta(\alpha)$. Using the identity

$$|H(z) - \sigma_\alpha(z)| = \begin{cases} 1 - \sigma_\alpha(z) = \dfrac{1}{1+e^{\alpha z}}, & z \ge 0, \\ \sigma_\alpha(z) = \dfrac{1}{1+e^{-\alpha z}}, & z < 0, \end{cases} \tag{4.12}$$

we obtain the symmetric form

$$|H(z) - \sigma_\alpha(z)| = \frac{1}{1+e^{\alpha|z|}} \le e^{-\alpha|z|}. \tag{4.13}$$

A voxel-wise application yields

$$\delta(\alpha) \le \frac{1}{N}\sum_{i=1}^{N} e^{-\alpha|z_i|}. \tag{4.14}$$

To obtain an interpretable and clinically meaningful bound, let q_m denote the fraction of voxels whose dose values lie within a margin m around the threshold:

$$q_m = \frac{1}{N}\#\{i : |z_i| \le m\}. \tag{4.15}$$

For voxels inside this margin, the worst-case error is 1/2 (since $\sigma_\alpha(0) = 1/2$). For voxels outside the margin, the error is bounded by $e^{-\alpha m}$. Thus both $\delta(\alpha)$ and the approximation error of $V_{x\%}$ in Eq. (4.11) satisfy

$$\left|V_{x\%} - V_{x\%}^{\text{approx}}(\alpha)\right| \le \delta(\alpha) \le q_m\left(\tfrac{1}{2}\right) + (1 - q_m)e^{-\alpha m}. \tag{4.16}$$

To ensure that the total approximation error remains below a user-specified tolerance ε (%), it suffices that

$$q_m\left(\tfrac{1}{2}\right) + (1 - q_m)e^{-\alpha m} \le \varepsilon. \tag{4.17}$$

Solving for α gives the required lower bound

$$\alpha \ge \frac{1}{m}\ln\left(\frac{1 - q_m}{\varepsilon - \frac{1}{2}q_m}\right). \tag{4.18}$$

Eq. (4.18) provides a lower bound on the slope parameter, ensuring a balance between approximation accuracy and avoidance of gradient saturation.

In summary, for each *V–metric* we first choose a tolerance ε that serves as an upper bound on the error tolerated by the user. We then define a margin m around the dose threshold and compute q_m, the fraction of voxels whose dose values fall within this margin, from the dose distributions of the training set. Finally, we take the smallest α satisfying Eq. (4.18) and use it in the loss calculation.

4.3.1.4 Overall loss formulation

Using the differentiable formulation of *D–metrics* and the differentiable surrogate of *V–metrics* described above, we compute each metric in the template (Table 4.2) for both the predicted and ground truth dose in the same way. The loss is then defined as:

$$\mathcal{L}_{\text{CDM}} = \sum_{r\in\mathcal{R}}\sum_{k\in\mathcal{M}_r} w_{r,k}\left|M_{r,k}^{\text{pred}} - M_{r,k}^{\text{gt}}\right|, \tag{4.19}$$

where $M_{r,k}^{\text{pred}}$ and $M_{r,k}^{\text{gt}}$ denote the predicted and ground truth metric k for ROI r, and $w_{r,k}$ denotes a weight. To preserve voxel-wise accuracy of the predicted dose distribution, we also use the MAE loss:

$$\mathcal{L}_{\text{MAE}} = \frac{1}{|\Omega|}\sum_{v\in\Omega}\left|D^{\text{pred}}(v) - D^{\text{gt}}(v)\right|. \tag{4.20}$$

The final objective combines both components:

$$\mathcal{L}_{\text{total}} = \lambda_1\mathcal{L}_{\text{MAE}} + \lambda_2\mathcal{L}_{\text{CDM}}, \tag{4.21}$$

where λ_1 and λ_2 balance voxel-wise consistency and alignment of clinical objectives.

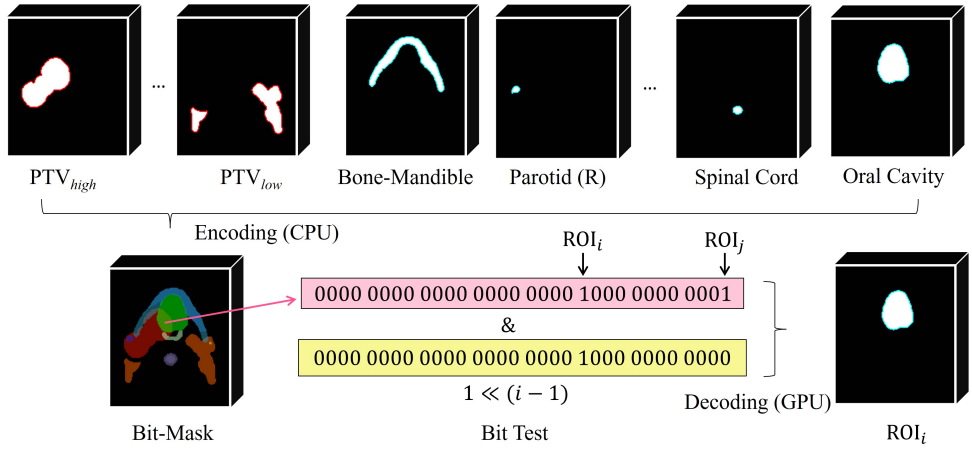


Figure 4.3: Multiple individual ROI masks are losslessly encoded into a single integer bit-mask, where each bit position represents a specific ROI (illustrated in pink). Because the raw bit-mask values span a wide numerical range, direct visualization is not intuitive. Therefore, the bit-mask is visualized in pseudo-color, with colors assigned solely based on their unique values for display purposes. During decoding, the binary mask for ROI$_i$ is obtained by testing whether the $(i-1)$-th bit of the bit-mask is active, implemented as a bitwise AND operation on GPU between the bit-mask and $(1 \ll (i-1))$, where the left-shifted single-bit mask $(1 \ll (i-1))$ is illustrated in yellow. A nonzero result indicates voxel membership in ROI$_i$.

4.3.2 Bit-mask encoding

The CDM loss in Chapter 4.3.1 requires access to all ROIs listed in Table 4.2. In H&N RT, the number of ROIs is typically large (often exceeding twenty), and many of them overlap spatially. In conventional dose prediction pipelines, each ROI is stored as an independent 3D binary mask channel using a one–hot encoding. When full 3D volumes are used as network input, this representation leads to substantial overhead in CPU preprocessing and CPU—to—GPU data transfer.

To alleviate these inefficiencies, we introduce a lossless bit–mask encoding as a compact intermediate representation for ROI handling, as illustrated in Fig. 4.3. Importantly, the bit-mask is not used as the direct input to the network. Instead, it serves as an efficient representation where standard binary ROI masks are decoded when needed, either to form the network input or to compute ROI-specific loss terms.

Formally, let $S_i(\mathbf{r}) \in \{0,1\}$ denote voxel membership in ROI$_i$. Assigning one bit position to each ROI yields the lossless encoding

$$B(\mathbf{r}) = \sum_{i=1}^{N_{\text{ROI}}} S_i(\mathbf{r})\, 2^{\,i-1}, \qquad N_{\text{ROI}} \le B_{\max}, \tag{4.22}$$

where $B_{\max}$ denotes the bit width of the integer type. In this work, a 32-bit unsigned

integer (`uint32`) is used, supporting up to 32 ROIs. Overlapping anatomical structures are represented by activating multiple bits within the same voxel, preserving all ROI relationships in a single channel.

A key advantage of this representation arises during data preprocessing. With a conventional one–hot encoding, geometric augmentations such as rotations, flips, and affine transformations must be applied independently to each ROI channel, causing preprocessing cost to scale linearly with the number of ROIs. In contrast, when ROIs are stored in a single bit-mask volume, each augmentation operation is applied once to the bit-mask, implicitly transforming all ROIs simultaneously. This substantially reduces CPU preprocessing time while maintaining exact ROI geometry.

The bit-mask representation also reduces CPU—to—GPU data transfer. Instead of transferring the CT volume together with more than twenty one–hot ROI channels, only the CT volume and a single compact bit-mask are transferred per iteration. This reduces I/O overhead and improves training throughput, especially when using full-resolution 3D volumes.

Once on the GPU, the bit–mask B is decoded into binary ROI masks as required. When forming the network input, the decoded ROI masks $\hat{S} \in \mathbb{R}^{N_{\text{ROI}} \times D \times H \times W}$ are concatenated with the CT volume to obtain

$$I_{\text{in}} \in \mathbb{R}^{(1+N_{\text{ROI}}) \times D \times H \times W}. \tag{4.23}$$

Thus, the network always operates on standard one–hot ROI channels, identical to those used in conventional pipelines. The bit-mask itself serves only as an intermediate encoding and does not alter the inputs of the network.

The decoding is implemented via lightweight, element-wise bit-test, as shown in Fig. 4.3. In particular, the binary mask for ROI_i is obtained via

$$S_i(\mathbf{r}) = \text{bool}(B(\mathbf{r}) \,\&\, (1 \ll (i-1))), \tag{4.24}$$

where $\text{bool}(\cdot)$ maps nonzero values to 1, "&" denotes the bitwise AND, and "$\ll$" denotes the bit left-shift operation. These operations are highly efficient on modern GPUs and introduce negligible computational overhead.

During loss computation, only the bit–mask B remains resident in GPU memory, and the binary mask for each ROI_i is decoded on demand at the moment its corresponding metric $M^{\text{pred}}_{r,k}$ or $M^{\text{gt}}_{r,k}$ is evaluated. The decoded mask is used to gather voxel doses for that ROI and is then immediately released. This on–demand decoding strategy avoids storing all N_{ROI} ROI channels simultaneously and thereby reduces peak GPU memory usage.

Table 4.3: Comparison of model performance under different loss-function configurations. Lower scores indicate better performance.

Loss function	PTV Score (%) ↓	OAR Score (Gy) ↓	Dose Score (Gy) ↓
MAE	1.544 ± 1.188	2.103 ± 0.704	**1.300 ± 0.229**
MAE + DVH	0.992 ± 0.466	2.162 ± 0.518	1.385 ± 0.217
MAE + DVH + CDM	0.567 ± 0.300	**1.933 ± 0.481**	1.389 ± 0.228
MAE + CDM (proposed)	**0.491 ± 0.250**	1.999 ± 0.497	1.370 ± 0.245

4.4 Experiments and results

4.4.1 Experiment setup

The ROIs used in our experiments are listed in Table 4.2. The template contains two PTV structures ($PTV_{54.25}$ and PTV_{70}) and 28 OARs, resulting in a total of $N_{ROI} = 30$ structures. Because this number fits within a 32-bit range, all ROI masks are encoded using a 32-bit unsigned integer (`uint32`) following the bit–mask representation described in Chapter 4.3.2.

For the CDM loss, all PTV-related $V_{95\%}$ metrics were included in the differentiable surrogate of *V–metrics*. To ensure controlled approximation accuracy, we adopt a margin of $m = 0.5\,\text{Gy}$ (i.e., a 1 Gy window) and a tolerance of $\varepsilon = 1\%$. Using the lower-bound condition in Eq. (4.18), we compute the minimum α values that satisfy the prescribed tolerance. Thus, we set $\alpha = 209$ for $PTV_{54.25}$ and $\alpha = 176$ for PTV_{70}. To reflect the higher clinical priority of PTVs, all PTV-related metric weights are set to $w(r,k) = 1$, whereas all OAR-related metric weights are set to $w(r,k) = 0.1$. The balancing coefficients are set to $\lambda_1 = 1$ and $\lambda_2 = 0.5$ for the final loss $\mathcal{L}_{\text{total}}$.

Unless otherwise stated, all experiments were conducted using a 3D U-Net baseline implemented in the `MONAI`[2] framework. The architecture follows the standard encoder–decoder design with skip connections [106]. Compared with the conventional 3D U-Net, the major architectural modification is the use of 3D Instance Normalization instead of Batch Normalization, which has been shown to provide more stable training for small-batch 3D medical imaging DL tasks [107].

All models were trained for 1000 epochs using the AdamW optimizer with an initial learning rate of 3×10^{-4} and a cosine annealing schedule. Mixed-precision training (bf16-mixed) was used throughout to reduce memory consumption and improve computational efficiency. All experiments were performed on an NVIDIA RTX 6000 Ada GPU.

4.4.2 Results

4.4.2.1 Loss function comparison

We first compared different loss-function configurations using the 3D U-Net baseline to evaluate the impact of the proposed CDM loss. Four loss-function configurations were investigated: MAE, MAE + DVH, MAE + DVH + CDM, and MAE + CDM (our proposed setting). Here, the DVH loss refers to the DVH value loss introduced in [26], which provides a more efficient way to penalize DVH-curve discrepancies compared with the formulation in [95]. In all experiments, the MAE loss was weighted as 1, while both the DVH and CDM components were assigned weights of 0.5. Table 4.3 reports the model performance under different loss-function configurations. The results show that introducing the CDM loss led to a substantial improvement in the PTV Score, achieving the lowest value when combined with MAE. In contrast, both MAE and MAE + DVH alone failed to effectively reduce the PTV Score. The CDM loss also provided moderate benefits for the OAR Score, suggesting its positive influence on balancing target coverage and organ-at-risk sparing. Although the MAE-only configuration yielded the lowest Dose Score, it also had a very high PTV Score.

We further examined whether the predicted dose distributions satisfied the predefined planning constraints. As illustrated in Fig. 4.4, the MAE + CDM configuration was the only setting that met both the $V_{95\%}$ ($\geq 98\%$) and $D_{0.03cc}$ ($\leq 110\%$) constraints across all test cases. No statistically significant differences from the clinical plans were observed for these metrics. In contrast, the other loss configurations showed instances of insufficient PTV coverage or excessive dose, as shown in Fig. 4.5. For OARs, all loss configurations satisfied the relevant clinical limits, but only the MAE + CDM model showed no statistically significant differences from the clinical plans for all metrics.

4.4.2.2 Ablation study

To assess the influence of α used in the differentiable *V–metric* approximation, we performed an ablation study on the parameter α (Table 4.4). Smaller values (e.g., 100) resulted in higher PTV Scores. The theoretically derived values $(209, 176)$ for $\text{PTV}_{54.25}$ and PTV_{70} achieved the best PTV Score while maintaining stable OAR and Dose Scores. Increasing α beyond this range (300–500) did not yield further improvement and in some cases slightly degraded PTV performance.

To further assess the computational impact of the proposed bit–mask ROI encoding, an ablation study was performed using the same 3D U-Net baseline and identical training settings. The comparison is reported in Table 4.5. When using conventional one–hot ROI channels, training required a peak GPU memory of 20.50 GB and an average epoch duration of 241 s. With bit–mask encoding, the peak memory was

[2] `https://github.com/Project-MONAI/MONAI`

Table 4.4: Ablation study on the parameter α. Lower scores indicate better performance.

α ($PTV_{54.25}$)	α (PTV_{70})	PTV Score (%) ↓	OAR Score (Gy) ↓	Dose Score (Gy) ↓
100	100	0.564 ± 0.287	**1.919 ± 0.502**	1.349 ± 0.228
209	176	**0.491 ± 0.250**	1.999 ± 0.497	1.370 ± 0.245
300	300	0.500 ± 0.276	1.961 ± 0.491	**1.346 ± 0.226**
400	400	0.563 ± 0.298	1.988 ± 0.475	1.352 ± 0.214
500	500	0.563 ± 0.308	1.993 ± 0.452	1.327 ± 0.204

Table 4.5: Ablation experiment on the 3D U-Net baseline comparing computational efficiency with and without the bit–mask ROI encoding.

Method (3D U-Net baseline)	Peak GPU Memory (GB)	Epoch Duration (s)
Without bit–mask	20.50	241
With bit–mask (proposed)	**19.57**	**43**

Table 4.6: Comparison of dose prediction performance across different network architectures using the MAE + CDM loss configuration. Lower scores indicate better performance.

Model	PTV Score (%) ↓	OAR Score (Gy) ↓	Dose Score (Gy) ↓
SwinUNETR [45]	0.699 ± 0.292	2.035 ± 0.566	1.461 ± 0.253
C3D [91]	0.497 ± 0.314	**1.945 ± 0.473**	**1.289 ± 0.192**
MedNeXt [108]	0.496 ± 0.251	2.154 ± 0.571	1.395 ± 0.204
Pyfer [24]	**0.491 ± 0.287**	2.019 ± 0.517	1.290 ± 0.207
3D U-Net (baseline) [106]	**0.491 ± 0.250**	1.999 ± 0.497	1.370 ± 0.245

19.57 GB (4.5% reduction) and the epoch duration was 43 s (82.2% reduction). Peak GPU memory is reported because memory usage during training is dynamic, and the maximum allocation determines whether the model fits within available GPU memory.

4.4.2.3 Model comparison

Building upon the best-performing loss setting (MAE + CDM), finally we compared the 3D U-Net baseline with several state-of-the-art architectures (Table 4.6), including transformer-based models (SwinUNETR [45]), large-kernel ConvNeXt-style convolutional networks (MedNeXt [108]), and cascade-style frameworks (C3D [91], Pyfer [24]). Notably, the 3D U-Net achieved a PTV Score of 0.491, matching the best-performing model (Pyfer) and outperforming more complex architectures such as SwinUNETR and MedNeXt. The OAR Score of 1.999 was also competitive, ranking second among all models.

4.5 Discussion

Overall, our results demonstrate that directly optimizing clinically used dose metrics is substantially more effective than voxel-wise or DVH-curve–based supervision. As shown in Table 4.3, MAE alone achieved the lowest voxel-wise Dose Score, but failed to ensure adequate target coverage. Previously proposed DVH-based losses improved DVH curve similarity [95, 104]. However, they did not reliably enforce clinically mandatory constraints, such as $V_{95\%}$ for PTVs. This limitation is evident in Fig. 4.4 and Fig. 4.5. Both MAE and MAE + DVH exhibit cases of insufficient target coverage or excessive hot spots. In contrast, the proposed CDM loss directly optimizes the dose metrics used in clinical plan evaluation. This task-aligned supervision consistently satisfied all PTV constraints in all patients, without compromising OAR sparing. These results indicate that the proposed CDM loss improves clinical relevance and outperforms loss formulations based on voxel-wise or DVH-curve supervision.

A key component of the proposed framework is the differentiable surrogate for *V–metrics*, which uses a sigmoid slope parameter α. The analytically derived lower bound provides a principled and reproducible way to select α without requiring exhaustive tuning. The ablation study (Table 4.4) shows that a small α leads to overly smooth threshold approximations. As a result, the model struggled to distinguish doses above and below the clinical prescription dose level. In contrast, very large α values created overly sharp transitions. This limits the number of voxels that contribute meaningful gradients and makes it harder for the model to correct small deviations near the prescription threshold. Between these extremes, the analytically derived values (209, 176) in this study provided stable performance. Additionally, the results also show that slightly larger values such as $\alpha = 300$ achieve similar PTV, OAR, and Dose Scores. Accordingly, the chosen values should be interpreted as a conservative and principled choice, rather than an attempt to identify an optimal setting.

An important advantage of the proposed CDM loss is its flexible, template-driven design. All optimized metrics, ROIs, and their relative priorities are specified through a simple JSON-based clinical evaluation template. This allows the loss to be configured for different planning protocols without modifying the network architecture or training pipeline. Although this study focuses on a single-institution H&N dataset, the template-based formulation in principle can be adapted to other datasets by updating the clinical criteria template. Compared with fixed, hard-coded loss formulations, this approach provides a practical and transparent way to align training objectives with clinically used evaluation standards.

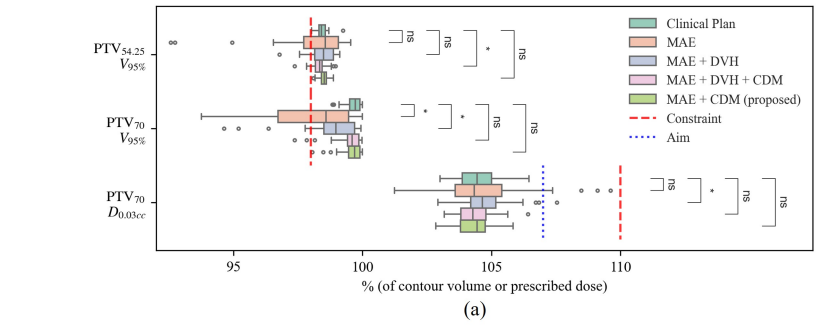
Beyond loss formulation, this work addresses a practical but often overlooked bottleneck in H&N dose prediction: the computational overhead caused by a large number of ROIs. The proposed bit-mask encoding strategy replaces more than twenty

one–hot ROI channels with a single integer-valued mask and decodes ROI membership on the GPU. This substantially reduces training time, with the average epoch duration decreasing from 241 s to 43 s (Table 4.5), accelerating training by more than 5×. In contrast to conventional pipelines [22, 91, 24, 28] that often restrict supervision to a subset of structures due to computational constraints, the proposed encoding enables more comprehensive and clinically faithful supervision. This supports more informative learning of spatial dose relationships across PTVs and OARs.

An additional observation from the model comparison experiments is that architectural complexity becomes a secondary factor when the training objective is well aligned with clinical goals. Under the MAE + CDM loss, a standard 3D U-Net achieved performance comparable to, or better than, transformer-based [45] and cascade-style architectures [91, 24] (Table 4.6). This suggests that, for 3D dose prediction, improving *what the model is optimized for* may be more impactful than increasing architectural complexity. From a clinical deployment perspective, this is an encouraging result, as simpler architectures are typically easier to train, validate, and maintain.

Several limitations of this study should be acknowledged. First, all data were obtained from a single institution with a unified planning protocol. The generalizability of the approach to other institutions therefore remains to be validated. Second, while the predicted dose distributions satisfied all clinical constraints, this study did not evaluate downstream integration into deliverable treatment planning [85, 109]. Future work will focus on multi-institutional validation and on assessing the impact of CDM-guided dose prediction on dose mimicking and final plan quality.

In summary, this work presents a flexible and clinically aligned framework for 3D head and neck dose prediction that combines a clinical DVH metric loss with an efficient, lossless ROI encoding strategy. By directly optimizing clinically relevant evaluation metrics, the proposed approach satisfied all predefined clinical constraints and improves the clinical acceptability of the predicted dose distributions. As such, the framework provides a practical foundation for AI-assisted radiotherapy workflows and may facilitate automated planning by improving plan consistency and reducing manual effort.



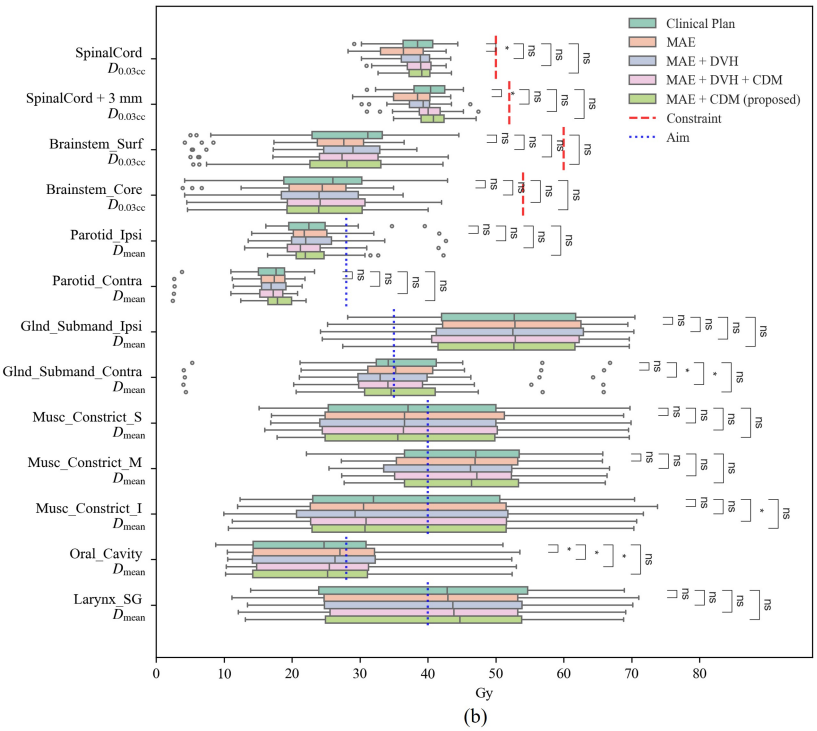


Figure 4.4: Boxplots of dose evaluation metrics for the clinically most relevant ROIs for different loss function configurations. (a) Boxplots for PTVs. (b) Boxplots for OARs. The corresponding planning aims and dose constraints are also shown (aims are clinically regularly violated due to proximity of the tumor). For $V_{95\%}$, the constraint is defined as ≥ the corresponding threshold, whereas for D_{mean} and $D_{0.03\text{cc}}$, both the constraint and the aim are defined as ≤ the corresponding threshold. Parotid and submandibular glands were categorized as ipsilateral (Ipsi) or contralateral (Contra) according to received dose. Statistical significance was tested using a two-tailed Wilcoxon signed-rank test. *: $p \leq 0.05$, ns = not significant.

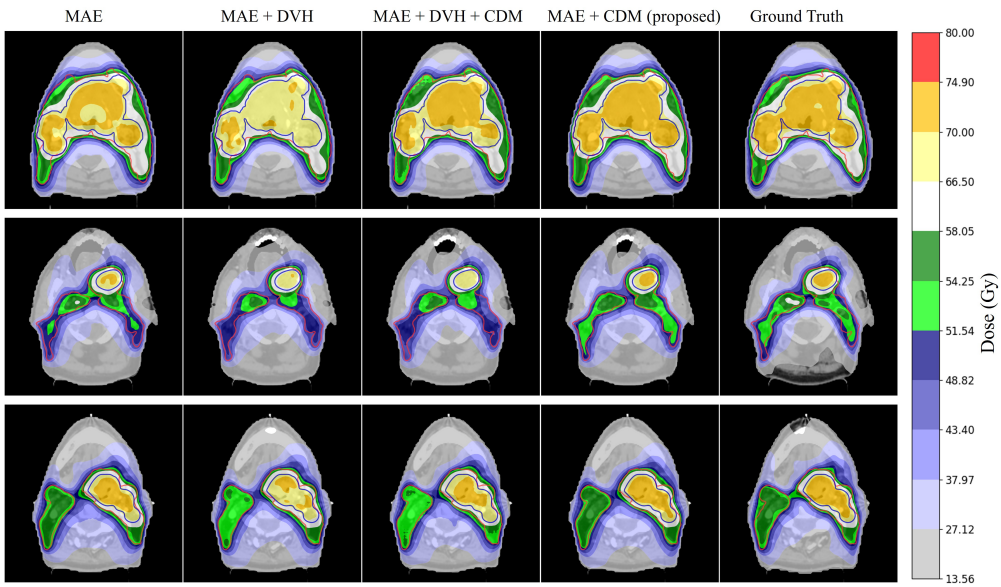


Figure 4.5: Axial visualization of dose distributions under different loss function configurations. The red contour denotes the $PTV_{54.25}$, while the blue contour represents the PTV_{70}. A colormap routinely used in clinical practice is employed to enhance visual interpretability.