Physics in Medicine & Biology

IPEM
Institute of Physics and Engineering in Medicine

PAPER

An ultra-fast deep-learning-based dose engine for prostate VMAT via knowledge distillation framework with limited patient data

To cite this article: Wenchih Tseng *et al* 2023 *Phys. Med. Biol.* **68** 015002

IOP Publishing Phys. Med. Biol. 68 (2023) 015002 https://doi.org/10.1088/1361-6560/aca5eb

Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

RECEIVED
29 August 2022

REVISED
16 November 2022

ACCEPTED FOR PUBLICATION
24 November 2022

PUBLISHED
19 December 2022

PAPER

An ultra-fast deep-learning-based dose engine for prostate VMAT via knowledge distillation framework with limited patient data

Wenchih Tseng[1], **Hongcheng Liu**[2,*], **Yu Yang**[2], **Chihray Liu**[1] and **Bo Lu**[1,*]

[1] Department of Radiation Oncology, University of Florida, Gainesville, FL 32610-0385, United States of America
[2] Department of Industrial and Systems Engineering, University of Florida, Gainesville, FL 32611-6595, United States of America
[*] Authors to whom any correspondence should be addressed

E-mail: hliu@ise.ufl.edu and lubo@shands.ufl.edu

Keywords: dose calculation engine, VMAT, knowledge distillation, deep learning

Abstract

Objective. Deep-learning (DL)-based dose engines have been developed to alleviate the intrinsic compromise between the calculation accuracy and efficiency of the traditional dose calculation algorithms. However, current DL-based engines typically possess high computational complexity and require powerful computing devices. Therefore, to mitigate their computational burdens and broaden their applicability to a clinical setting where resource-limited devices are available, we proposed a compact dose engine via knowledge distillation (KD) framework that offers an ultra-fast calculation speed with high accuracy for prostate Volumetric Modulated Arc Therapy (VMAT). *Approach*. The KD framework contains two sub-models: a large pre-trained teacher and a small to-be-trained student. The student receives knowledge transferred from the teacher for better generalization. The trained student serves as the final engine for dose calculation. The model input is patient computed tomography and VMAT dose in water, and the output is DL-calculated patient dose. The ground-truth \dose was computed by the Monte Carlo module of the Monaco treatment planning system. Twenty and ten prostate cases were included for model training and assessment, respectively. The model's performance (teacher/student/student-only) was evaluated by Gamma analysis and inference efficiency. *Main results*. The dosimetric comparisons (input/DL-calculated/ground-truth doses) suggest that the proposed engine can effectively convert low-accuracy doses in water to high-accuracy patient doses. The Gamma passing rate (2%/2 mm, 10% threshold) between the DL-calculated and ground-truth doses was 98.64 ± 0.62% (teacher), 98.13 ± 0.76% (student), and 96.95 ± 1.02% (student-only). The inference time was 16 milliseconds (teacher) and 11 milliseconds (student/student-only) using a graphics processing unit device, while it was 936 milliseconds (teacher) and 374 milliseconds (student/student-only) using a central processing unit device. *Significance*. With the KD framework, a compact dose engine can achieve comparable accuracy to that of a larger one. Its compact size reduces the computational burdens and computing device requirements, and thus such an engine can be more clinically applicable.

1. Introduction

Dose calculation is an essential part of treatment planning in radiotherapy. Its calculation accuracy and efficiency can fundamentally affect the plan quality and planning time (Shepard *et al* 2002) especially when iterative dose calculation processes are conducted during optimization for inverse planning (e.g. Intensity Modulated Radiotherapy (IMRT), Volumetric Modulated Arc Therapy (VMAT) planning). However, a highly accurate dose calculation algorithm, e.g. Monte Carlo (MC) simulation, typically requires a great deal of computational power (Chen *et al* 2014b, Xing *et al* 2020a, 2020b). Therefore, to improve planning efficiency while maintaining a high level of accuracy for the final computed dose, commercialized treatment planning systems (TPSs), such as Pinnacle (Philips Medical Systems, Madison, WI), Monaco (Elekta AB, Stockholm, Sweden), etc, typically adopt

a low accuracy (yet fast) dose calculation algorithm (e.g. pencil beam algorithms) from the first iteration up until the last iteration, or the last couple of iterations, of the optimization, at which point a highly accurate (yet slow) dose calculation algorithm (e.g. Superposition/Convolution (S/C), MC simulation) is applied (Li *et al* 2015). This may compromise treatment plan quality, resulting in a sub-optimal plan.

To overcome these challenges, a dose engine should ideally possess not only a high level of efficiency but also a high level of accuracy for inverse planning. In recent years, many deep learning (DL) methods (Peng *et al* 2019, Fu *et al* 2020, Kontaxis *et al* 2020, Xing *et al* 2020a, 2020b, Liu *et al* 2021, Tsekas *et al* 2021) have been proposed for such dose engines. Two major strategies have been implemented for DL-based dose engines. One of the strategies (Kontaxis *et al* 2020, Tsekas *et al* 2021) encodes the beam information into the learning phase of a DL model. (Kontaxis *et al* 2020) proposed a so-called 'Deep Dose' framework, which incorporates the treatment information into the patient anatomy for segment dose calculation of five-field (fixed) IMRT plans, using data from 101 prostate patients (4176 segments). It can achieve an average Gamma passing rate of 97.8% and 99.5% using 2%/2 mm Gamma criterion with a 10% dose threshold for segment dose and plan dose, respectively. The average calculation time was 0.6 s per segment and 25 s per plan on an NVIDIA GTX Titan graphics processing unit (GPU). This framework was also found to be effective in IMRT dose calculation within a 1.5 T magnetic field (Tsekas *et al* 2021). The other strategy (Peng *et al* 2019, Fu *et al* 2020, Xing *et al* 2020a, 2020b, Liu *et al* 2021) turns the dose calculation task into a dose conversion problem (from low-accuracy doses to high-accuracy ones) using a DL model as a de-noising tool. Xing *et al* (2020a) trained a DL-based dose engine for 7-field (fixed) IMRT plans using doses acquired from an inaccurate (but fast and low-cost) algorithm (Lu and Chen 2010) and a highly-accurate algorithm (S/C algorithm (Ahnesjö 1989)) as the input and reference outputs, respectively. Data for a total of 78 prostate patients were collected for engine training. The model can achieve an average Gamma passing rate of 98.5% (1%/1 mm) and 99.9% (2%/2 mm) at a 20% dose threshold. The average total calculation time on an NVIDIA Tesla V100 GPU was 1.19 s per plan. Three other research groups (Peng *et al* 2019, Fu *et al* 2020, Bai *et al* 2021) used the same strategy to de-noise MC doses that possessed a high statistical uncertainty to obtain ones with a lower statistical uncertainty to accelerate the planning dose calculation.

The aforementioned research groups (Peng *et al* 2019, Fu *et al* 2020, Kontaxis *et al* 2020, Xing *et al* 2020a, 2020b, Bai *et al* 2021, Liu *et al* 2021, Tsekas *et al* 2021) have demonstrated that a DL-based dose engine is able to offer superior computational speed compared to traditional high-accuracy algorithms, while maintaining a comparable level of accuracy. However, much room for improvement remains for the existing framework of DL-based dose engines to broaden their clinical applicability. First, most of the DL-based dose calculation algorithms were specifically designed for fixed-field IMRT plans (Peng *et al* 2019, Kontaxis *et al* 2020, Xing *et al* 2020a, Liu *et al* 2021, Tsekas *et al* 2021). The same calculation speed and accuracy cannot be assumed for VMAT dose calculation. Compared to the fixed-field (fixed gantry angle) IMRT plans, a more complex beam modulation is generally expected for VMAT plans as its dynamic nature with simultaneous changes of gantry rotational speed, dose rate, and multi-leaf collimator arrangements during beam delivery (Bedford 2009, Chen *et al* 2014a, Unkelbach *et al* 2015). Second, current DL models typically possess high computational complexity, and a large storage capacity is commonly needed for highly accurate dose calculations. This usually requires a powerful computing resource (e.g. GPU), which implies a high cost for computation. It is therefore challenging to deploy such a model on low-powered and resource-limited devices. Consequently, a small-capacity DL model with lower computational burdens seems to be more clinically applicable. Third, to improve the model performance on unseen data, a large and comprehensive patient database is typically included in the training phase, which necessitates a difficult data collection task, as accessible patient data are of limited availability in many clinical settings. Building a robust DL-based dose engine with limited patient data still remains a great challenge. Therefore, an approach to effectively augment the training dataset for the dose calculation task would be beneficial to the success of the DL-based dose engine.

In this study, we developed a novel DL-based dose engine for prostate VMAT plans to address the aforementioned limitations. First, instead of performing the dose calculation per aperture or control point (CP) of a VMAT arc, the proposed framework enables a composite arc dose calculation via the composite arc dose in water and the patient computed tomography (CT) images. This allows for a much faster VMAT dose calculation, and thus a shorter calculation time can be expected. Second, knowledge distillation (KD) (Hinton *et al* 2015), which is a model compression technique that distills the knowledge of a large-capacity model (teacher) into a smaller one (student) without a severe loss of soundness in performance, was implemented to further decrease the computational burdens of the DL model. As a result, the trained student can achieve a computational accuracy that is comparable to that of the teacher, but it is also able to attain a superior level of efficiency, which makes the model more applicable to a clinical setting of limited computational resources. Third, limited patient data with the arc recycling strategy were included for the model training (twenty patient cases) and performance assessment (ten patient cases). This strategy artificially inflates the size of the raw training dataset to mitigate the model overfitting without the requirement of extensive patient data collection. The detailed model implementations are presented in the Methods and Materials section. Comprehensive analysis of the model performance is

conducted and discussed. The potential clinical applications of the proposed dose engine are summarized at the end.

2. Methods and materials

In this section, we first introduce the patient CT and arc plan database for training in section 2.1. Subsequently, data generation and preprocessing (section 2.2) and the KD framework (section 2.3) are presented in detail. Finally, the approaches to measure the performance of the proposed DL-based dose engine are given in section 2.4.

2.1. Patient and arc database

Twenty prostate cancer cases were collected, of which sixteen and four were randomly selected as training and validation patients, respectively. Seventy VMAT arcs (6 MV full arcs with various collimator angles from 5° to 110°) were collected from these 20 cases. To enlarge the size of the dataset, each arc was recycled and applied to the different patient CTs for dose calculation, and so 1120 and 280 samples (involving one arc beam, one CT, and one computed dose matrix) were included in the training and validation datasets, respectively. This arc recycling strategy not only artificially increases the training dataset, it also facilitates the DL model with the learning of the effects of dosimetric variations due to various portions of the patient anatomy irradiated from the same beam, or due to different arc beams delivered to the same anatomical site. An additional ten prostate cases served as an independent testing dataset for performance evaluation. The detailed patient demographics and plan information were listed in table 1.

2.2. Data generation and preprocessing

Two-channel inputs were applied for model training as well as for final DL-based dose calculation. The patient CT images comprise the first channel of inputs. The Hounsfield Unit (HU) range of patient CT images is from 0 to 3000. The second channel is the three-dimensional (3D) dose for the VMAT plans, computed in a patient-contour-shaped water phantom. In contrast to other DL-based dose calculation studies (Peng *et al* 2019, Fu *et al* 2020, Xing *et al* 2020a, 2020b, Bai *et al* 2021) whose low-accuracy dose input was typically calculated by either MC simulations or other traditional dose calculation algorithms on real patient CTs, we performed a simple pencil beam convolution on a homogeneous (water) phantom (possessing the patient's external contour) to generate dose as the input. This method can be implemented easily and allows for a faster speed of dose input generation. On the other hand, the correlation of dose variation due to anatomical heterogeneity can be extracted with the information provided by the first-channel inputs in our model. That is, the patient CT images provide anatomical information to assist in mapping the doses from water to real patient anatomy with heterogeneity considerations.

To obtain the plan dose in the water phantom, we first compute the CP dose in the water phantom using the following pencil beam convolution formula:

$$D_i(d) = A_i(d) * K_i(d). \tag{1}$$

In gantry coordinates, $D_i(d)$ represents the 2D dose at depth d of the phantom, perpendicular to the beam's central axis direction, contributed by the ith CP. $A_i(d)$ is a 2D binary function, representing the beam projection of the ith CP aperture at depth d. It should be noted that the aperture projection region changes along the depth due to the beam divergence, as indicated in figure 1(a). $K_i(d)$ is the 2D pencil beam kernel of the ith CP at depth d. The nominal pencil beam kernel was simulated on a $50 \times 50 \times 50$ cm^3 water phantom with $0.1 \times 0.1 \times 0.1$ cm^3 resolution under an EGSnrc (Kawrakow 2000) MC environment. An enface 6 MV, 0.1×0.1 cm^2 beamlet (projected at source-to-axis distance, SAD) at 75 cm source-to-surface distance (SSD) was employed for the simulation. For the dose computation at each CP, the nominal kernel needs correction for beam divergence due to SSD variations using the inverse square law. The convolution process at different depths is illustrated in figure 1.

By computing the 2D dose for all depths (with 0.1 cm resolution), the 3D dose in gantry coordinates for the ith CP, namely D_i, can be subsequently acquired. Finally, the VMAT plan dose in the phantom in room coordinates can be found by summing all 3D CP doses using equation (2). Here, a rotation matrix $R(\theta_i)$ was applied to each CP dose D_i prior to final summation in order to convert the CP dose from gantry coordinates to room coordinates. An illustration of coordinate conversion is shown in figure 2

$$\text{VMAT dose} = \sum_{i=1}^{\text{CP}} R(\theta_i)\, D_i. \tag{2}$$

For each pair of two-channel inputs, the corresponding ground-truth plan dose was calculated on a Versa HD (Elekta Inc., Stockholm Sweden) machine model by the MC module of the Monaco TPS (version 5.51.02)

Table 1. Patient demographics and plan information used for the training/validation and testing datasets.

Dataset	Patient numbers	Tumor clinical stage	Prescription	Note
Training/Validation	20	Intermediate risk of prostate adenocarcinoma (stage II)	60 Gy	2 patients with titanium hip implant
Testing	10		(QD 3 Gy fraction^{-1})	1 patient with titanium hip implant

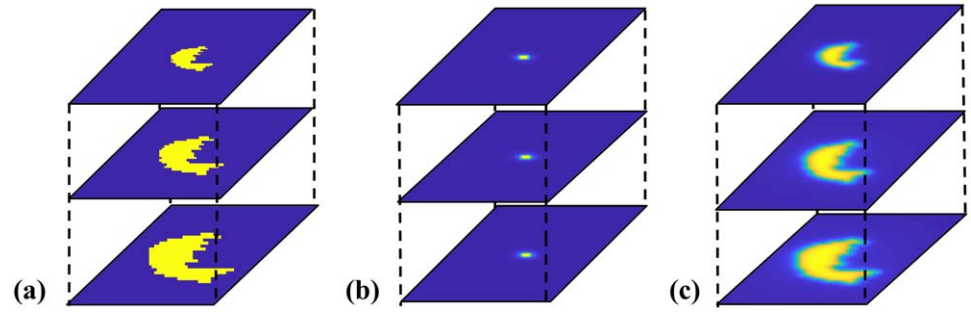

Figure 1. Visual illustrations of the (a) 3D binary CP beam projection, (b) 3D pencil beam dose, and (c) 3D CP dose in water.

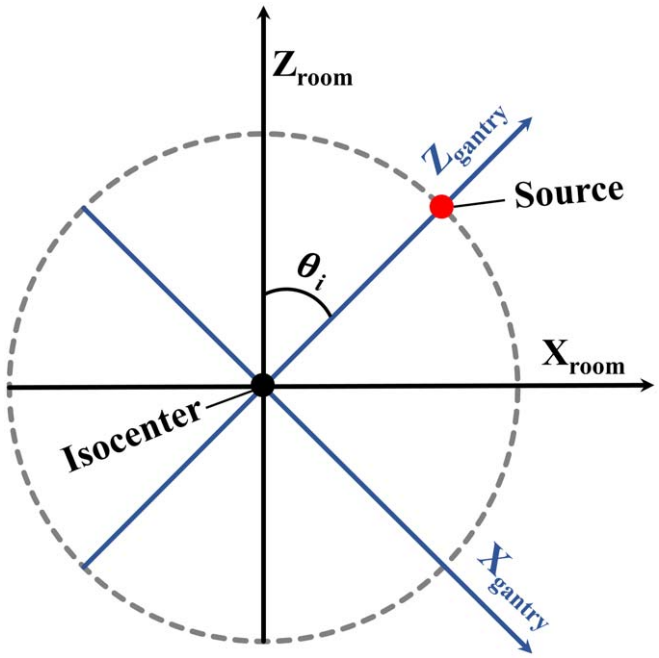

Figure 2. An illustration of coordinate conversion from gantry to room. θ_i is the rotational angle between gantry coordinates and room coordinates for the ith CP.

with a $0.3 \times 0.3 \times 0.3\ cm^3$ dose grid and a statistical uncertainty of 0.7% per calculation on a local computer equipped with dual 14 core Intel(R) 2.6 GHz Xeon(R) Gold 6132 central processing units (CPU) and 128 GB RAM. All input data including 3D doses, CT images, and ground-truth MC doses were cropped to the same size ($144 \times 96 \times 48$ voxels with a resolution of $0.3 \times 0.3 \times 0.3\ cm^3$) for the sake of simplicity. For training purposes, all dose values were normalized to planning monitor units (MUs).

2.3. KD framework

The KD framework (figure 3) is composed of two sub-models: a large pre-trained teacher model and a smaller to-be-trained student model. The student model is trained under the supervision of the teacher model, and mimics the teacher model's behaviors in order to achieve a competitive performance with that of the teacher model. In this framework, the training input and output to both teacher and student models are identical. The student model receives the transferred knowledge of the pre-trained teacher model by incorporating one extra distillation loss (the difference between the teacher's calculation and the student's calculation) into the loss calculation in the training phase. The densely connected neural network (Zhang *et al* 2020) was chosen as a base model architecture for both the teacher and student models' developments. The teacher model was trained first in order to provide good guidance for training a student model. We increased its number of trainable parameters by implementing different spatial resolution levels (2 to 4) and applying different initial numbers of feature channels (16–64) until its performance was no longer improved on. For the student model, we experimentally built a small-capacity model by limiting its number of trainable parameters. The trained

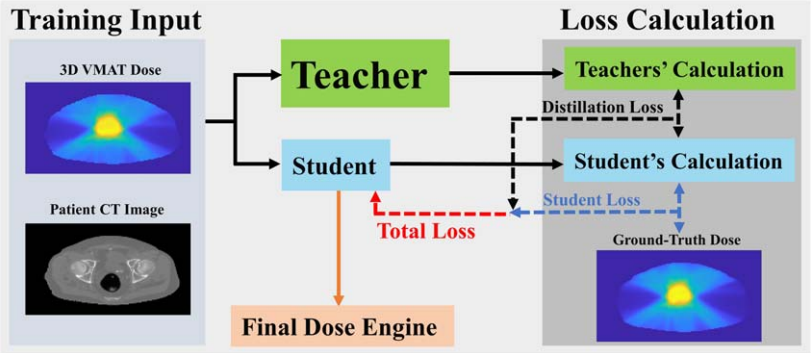

Figure 3. The teacher-student training framework for knowledge distillation.

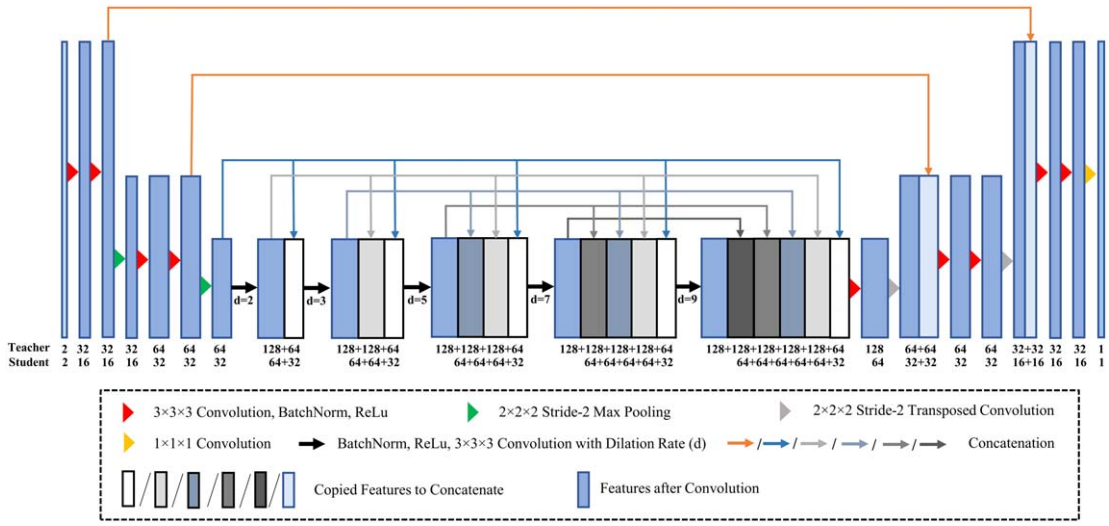

Figure 4. The architecture of the final teacher and student models. The number under each block represents the number of feature channels.

student model that has a performance comparable to that of the teacher model serves as the final dose engine for the prostate VMAT dose calculation. The detailed architecture of the final teacher and student model is described in section 2.3.1.

2.3.1. Teacher and student models' architecture

The architecture of the teacher and student models is illustrated in figure 4. It is composed of an encoding path and a decoding path. Each has three spatial resolution levels: $144 \times 96 \times 48$, $72 \times 48 \times 24$, and $36 \times 24 \times 12$. The first two levels in both paths consist of two convolutional layers with $3 \times 3 \times 3$ kernels and zero padding, followed by batch normalization and rectified linear units (ReLu). The down-sampling between levels in the encoding path is performed by a $2 \times 2 \times 2$ max pooling layer with a stride size of two, while the up-sampling between levels in the decoding path is conducted by a transposed convolutional layer with $3 \times 3 \times 3$ kernels and a stride size of two. The skip connection copies and concatenates the high-resolution features from the encoding path to the decoding path for preserving the local features. A dense feature aggregation block with five dilated convolutional layers (dilation rates: 2, 3, 5, 7, 9) is implemented at the bottleneck level, for which each convolutional layer connects to every other convolutional layer. The last convolutional layer with $1 \times 1 \times 1$ kernels reduces the feature channel to one for the final voxel-wise dose calculation. The feature channels are doubled as the spatial resolution is halved in the encoding path, whereas the feature channels are halved as the spatial resolution is doubled in the decoding path. The initial number of feature channel starts with 32 for the teacher model and 16 for the student model. Accordingly, the teacher model has trainable parameters of 6.3 million, while the student model has trainable parameters of 1.6 million.

2.3.2. Loss function

To train a teacher model, the mean absolute error (MAE) between the DL-calculated doses and the ground-truth MC doses was used as the loss function for optimization

$$\mathcal{L}_{\mathrm{MAE}} = \frac{1}{N}\sum_{n=1}^{N}|\mathrm{DL}_n - \mathrm{GT}_n|\,, \tag{3}$$

where $\mathcal{L}_{\mathrm{MAE}}$ is the MAE as a loss function. DL_n and GT_n are the DL-calculated dose and the ground-truth MC dose at the nth voxel, respectively. N is the total number of voxels in the 3D volume.

To incorporate the pre-trained teacher's knowledge into the training phase of the to-be-trained student model in the KD framework, the parameters of the student model were optimized by minimizing the total loss function as follows:

$$\mathcal{L}_{\mathrm{total}} = \alpha \cdot \mathcal{L}_{\mathrm{MAE}}(\mathrm{DL}_{\mathrm{student}}, \mathrm{GT}) + (1-\alpha)\cdot \mathcal{L}_{\mathrm{MAE}}(\mathrm{DL}_{\mathrm{student}}, \mathrm{DL}_{\mathrm{teacher}}), \tag{4}$$

where $\mathcal{L}_{\mathrm{total}}$ is the combined MAE of student loss and distillation loss. $\mathrm{DL}_{\mathrm{student}}$ and $\mathrm{DL}_{\mathrm{teacher}}$ are the DL-calculated doses from the student and teacher models, respectively. α is a weighting factor to combine the student loss and distillation loss, and it was set to 0.5 in this study. In the validation phase of the KD framework, the $\mathcal{L}_{\mathrm{MAE}}$ in equation (3) was used to evaluate the model performance on the validation dataset since only the output of the student model will be used for the final calculation.

2.3.3. Model training

The Adam algorithm with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$, and epsilon $= 10^{-8}$) and a learning rate of 0.0001 was selected as the optimizer to minimize the loss function. The training batch size and maximum epoch number were chosen to be 16 and 1000, respectively. An early stopping technique was implemented to terminate the training session at the point where the loss function did not improve by over 50 epochs. The model weights were randomly initialized. Five-fold cross-validation was first performed to assess the stability of the DL models and to search for the optimal hyperparameter setting. The final models were subsequently trained with the training and validation datasets combined (1400 arc samples in total), and then evaluated by the independent testing dataset. The aforementioned settings were applied to both the teacher and student models. The DL models were implemented with the PyTorch DL framework (version 1.10) and trained on a single NVIDIA A100 SXM4 GPU (80 GB RAM).

2.4. Performance assessment

To assess the performance of the proposed DL-based dose engine, two key evaluation metrics: (a) dosimetric accuracy and (b) inference efficiency, were mainly compared and discussed. First, the comparisons between the DL-calculated and ground-truth MC doses on the testing dataset were performed based on a global Gamma analysis using the criteria of both 2%/2 mm and 3%/3 mm with a low-dose threshold of 10%. The dosimetric performance of the proposed dose engine on tissue heterogeneity corrections was also evaluated using a prostate patient case with hip replacements and gas-filled rectum. Second, to measure the model inference efficiency on both powerful and limited computing devices, an NVIDIA A100 SXM4 GPU, an NVIDIA GTX 3080 GPU (10 GB RAM), and an Intel(R) Core (TM) 3.5 GHz i9–11900KF CPU (64 GB RAM) were utilized for dose calculation. Third, to study the benefits of the KD framework, the performance of a student-only model (trained without the KD framework) and the teacher model were also compared to the distilled student model.

3. Results

3.1. Model training results

The training session of the student model (final dose engine) trained with the KD framework took approximately 7 h on a single A100 GPU. Loss convergence curves of the training progress are presented in figure 5. Both training loss and validation loss decrease as the epoch increases. The training was stopped when the difference of validation losses between 50 consecutive epochs was lower than 0.0001, which ends up at 450 training epochs.

3.2. Dosimetric results of the final dose engine trained with the KD framework

3.2.1. Overall dosimetric performance

The detailed Gamma analysis of the test patients is reported in table 2. The mean as well as the standard deviation (SD) of the Gamma passing rates for the proposed dose engine on the test patients were 98.13 $\pm$ 0.76% and 99.50 $\pm$ 0.17% using 2%/2 mm and 3%/3 mm criteria, respectively. This indicates that the dose engine can produce highly accurate dosimetric results that are comparable to its counterpart calculated by the Monaco MC algorithm.

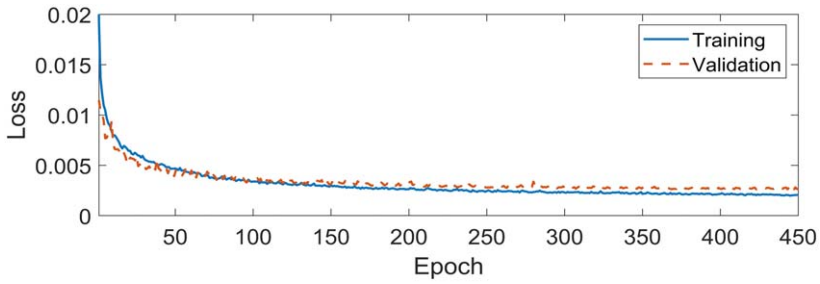

Figure 5. Training and validation loss convergence curves of the final dose engine.

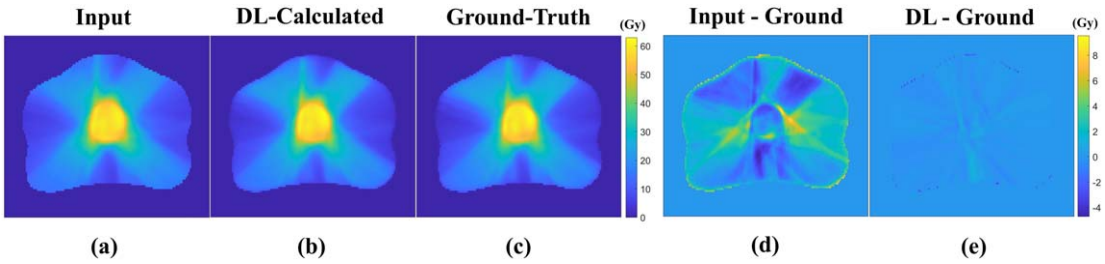

Figure 6. The central axial slice of the (a) input doses, (b) DL-calculated doses, and (c) ground-truth doses for the test patient 1. Absolute differences between ground-truth doses and either input doses or DL-calculated doses are presented in (d) and (e).

Table 2. The Gamma passing rates between the ground-truth doses and DL-calculated doses of the proposed dose engine on the ten test patients.

Patient index	Gamma passing rate (%)	
	2%/2 mm	3%/3 mm
1	99.07	99.65
2	99.27	99.71
3	98.66	99.52
4	98.04	99.54
5	97.45	99.43
6	97.11	99.20
7	97.48	99.31
8	98.69	99.67
9	97.47	99.38
10	98.10	99.54
Mean ± SD	98.13 ± 0.76	99.50 ± 0.17

Figure 6 provides visual comparisons of the dose distribution among the input doses, DL-calculated doses, and ground-truth doses for a sample case (test patient 1). Substantial differences between the input doses and ground-truth doses can be observed, compared to smaller differences between the DL-calculated doses and ground-truth doses. The Gamma passing rate between the input doses and ground-truth doses was 74.22% using a 2%/2 mm criterion with a 10% low-dose threshold, whereas the Gamma passing rate between the DL-calculated doses and ground-truth doses was 99.07%. This result suggests the learning capability of our model structure is strong, even with very coarse dose information as input.

Figure 7 shows the dose-volume histogram (DVH) of the PTV, and some organs-at-risk (OARs), calculated by both the MC method (ground-truth) and DL model for the sample case (patient 1). The DVH curves of the DL model are very close to their counterpart MC calculation for all of the structures.

3.2.2. Tissue heterogeneity corrections

Figure 8(a) provides a comparison of the isodose distributions between the DL-calculated doses and ground-truth doses for a sample case (test patient 4) with titanium hip implant and gas-filled rectum. The isodose lines of

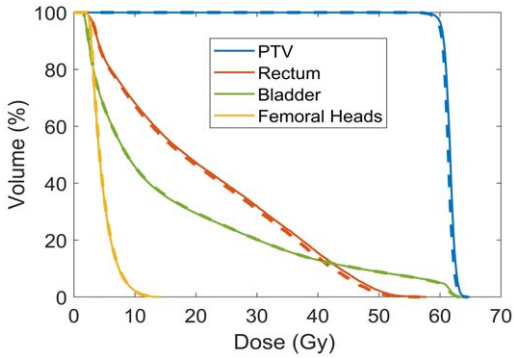

Figure 7. A DVH plot of the ground-truth (solid) and DL-calculated (dashed) dose distributions for the PTV, rectum, bladder, and femoral heads structures.

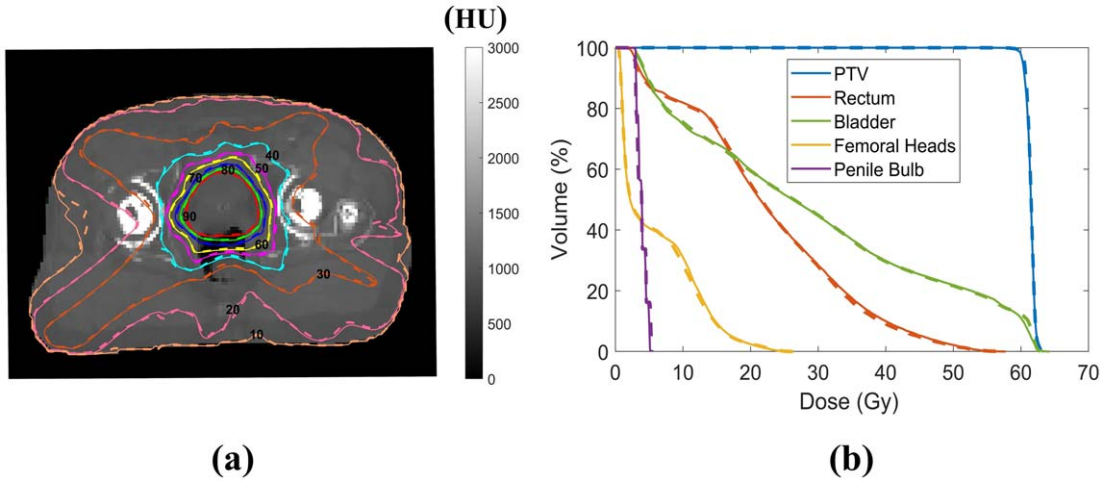

Figure 8. (a) A comparison of isodose distributions between the DL-calculated (solid) and ground-truth (dashed) doses for a patient case with titanium hip implant and gas filled rectum. (b) A DVH plot of the DL-calculated (solid) and ground-truth (dashed) dose distributions for the PTV and OARs structures.

the DL-calculated doses were matched well to those of the ground-truth doses. Similar agreement can also be observed in the DVH curves' comparisons as illustrated in figure 8(b). The Gamma passing rates for this case were 98.04% (2%/2 mm) and 99.54% (3%/3 mm). These results indicate that the proposed DL-based dose engine is able to properly handle the tissue heterogeneity corrections.

3.3. Dosimetric comparisons of the teacher, student, and student-only models

Figure 9 demonstrates the distributions of the Gamma passing rates on the test patients using the teacher model and the student-only model (trained without the KD framework), as well as the student model trained with the KD framework (final dose engine). The performance gap between a large-capacity model (teacher) and a small-capacity model (student-only) becomes smaller when such a small-capacity model (student) is trained with the KD framework. The Gamma passing rates on the ten test patients were $98.64 \pm 0.62\%$ for the teacher model and $96.95 \pm 1.02\%$ for the student-only model using a 2%/2 mm criterion, compared to $98.13 \pm 0.76\%$ for the student model. They were $99.63 \pm 0.16\%$ for the teacher model and $99.32 \pm 0.32\%$ for the student-only model using a 3%/3 mm criterion, compared to $99.50 \pm 0.17\%$ for the student model. These results suggest that the performance of a small-capacity DL model can be improved by the KD framework.

Figure 10 provides visual dose comparisons among the teacher, student, and student-only models using two sample cases (test patient 2 and test patient 3) on 2D planes. Noticeable variations between the isodose lines of the ground-truth doses and those from the student-only model can be observed. These differences (indicated by the black arrows) are moderately alleviated with the help of the KD framework, as can be seen in figure 10. The Gamma passing rates (2%/2 mm and 10% dose threshold) on test patient 2 were 99.46%, 99.27%, and 97.86%, for the teacher, student, and student-only models, respectively. They were 99.02%, 98.66%, and 97.69% on test patient 3.

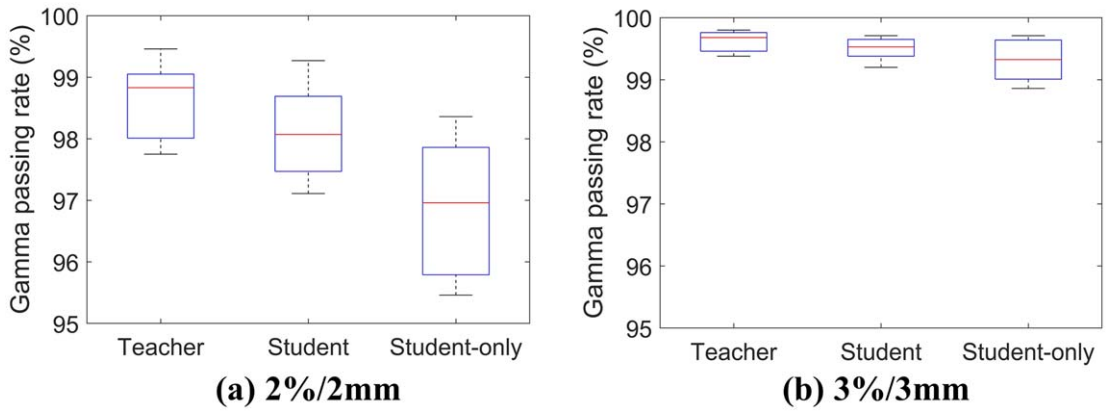

Figure 9. The distributions of the Gamma passing rates using (a) 2%/2 mm and (b) 3%/3 mm criteria on the ten test patients for the teacher, student, and student-only models.

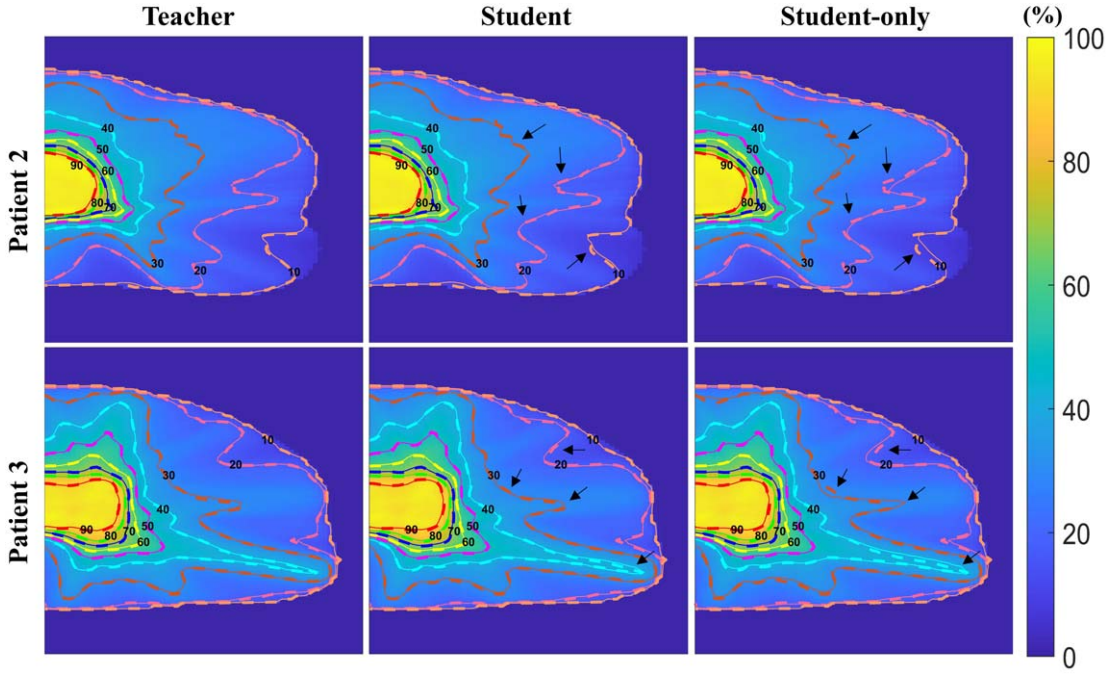

Figure 10. The dosimetric comparisons between the DL models' calculated doses (solid isodose lines) and ground-truth MC doses (dashed isodose lines) for two test patients. Doses were normalized to the maximum ground-truth dose. The number on the lines represents the percent dose. Planar dose maps were cropped to enlarge the details.

3.4. Model inference efficiency

Table 3 reports the model inference times for the DL-based dose calculations. The student and student-only models in general have the same level of inference efficiency since their model capacity is identical. Compared to the large-capacity teacher model, their smaller capacity allows for a shorter inference. An interesting observation is that the time-saving effect is actually improved for student models using the CPU, compared to that using the GPUs.

4. Discussion

4.1. The performance of the proposed DL-based dose engine

In this work, a small-capacity DL-based dose engine for prostate VMAT plans with limited patient data was presented. The KD framework was implemented to help improve the performance of the dose engine. In this way, this type of engine can offer an ultra-fast calculation inference while maintaining a high level of calculation accuracy. The results of the dosimetric comparisons (figures 6 and 7) show that the proposed dose engine can

Table 3. The inference time of the teacher, student, and student-only models. Batch size of 1 was used.

Model	Inference time per arc (unit: milliseconds)		
	GPU		CPU
	A100 312 TFLOPS[a]	RTX 3080 59.5 TFLOPS[a]	0.9 TFLOPS[a]
Teacher	16.0 ± 0.27	45.5 ± 0.57	936.2 ± 5.9
Student	11.0 ± 0.31	27.8 ± 0.28	374.8 ± 3.4
Student-only	10.8 ± 0.22	27.2 ± 0.29	± 2.8

[a] TFLOPS: a trillion floating-point operations per second used to measure the computational capacity of a computing device.

effectively translate low-accuracy arc doses in water into high-accuracy arc doses in patient anatomy for prostate VMAT plans. The dosimetric comparisons demonstrated in figure 8 further suggest that the proposed engine is capable of managing the tissue heterogeneity corrections in the presence of significant inhomogeneous prostate patient anatomy with high calculation accuracy. The statistical results of the Gamma analysis from table 2 demonstrate that the dose engine is able to compute with a high level of accuracy for all the test patients. In addition, the comparisons of the inference time reported in table 3 show that the proposed DL-based dose engine yields a much superior inference efficiency compared to that of the large-capacity model due to its compact size. These results suggest that the proposed dose engine is able to compute VMAT plans not only quickly, but accurately. It is therefore as promising tool for accelerating dose calculation and plan optimization processes for prostate VMAT planning.

4.2. The analysis of the proposed DL-based dose engine

To improve the calculation accuracy of the proposed small-capacity dose engine (final student model), the knowledge of the large-capacity teacher model was integrated into the training structure. The results of the isodose comparisons between the student model doses and the ground-truth doses (demonstrated in figure 10) have better agreement than those for the student-only model. Moreover, the distributions of the Gamma passing rates among the teacher, student, and student-only models (figure 9) indicate that the student model is able to generate the doses with better accuracy than student-only model and its computational accuracy closer to what the teacher model can achieve. A small gap between the Gamma passing rates of the teacher and student models is still noticeable. This discrepancy is expected due to the fact that the student model has a much smaller scale of parameters than the teacher model. Nevertheless, the proposed dose engine can still produce a level of dose accuracy that is comparable to some well-accepted algorithms, such as the C/S algorithm. According to our study, the Pinnacle TPS version 16.4 (C/S based algorithm) has a $98.26 \pm 0.85\%$ Gamma passing rate using a 2%/2 mm criterion on the test patients when compared to ground truth, which is very close to the passing rate that our model achieved, $98.13 \pm 0.76\%$.

From the calculation efficiency standpoint, our dose engine has a greater advantage over the MC simulation. For example, the calculation time per arc was approximately 3.3 s (input generation: 2.9 s and model inference: 0.4 s) and 180 s on CPU-only devices using the proposed dose engine and the Monaco MC module, respectively. In addition, comparing to other DL-based dose engines (Kontaxis *et al* 2020, Liu *et al* 2021, Tsekas *et al* 2021), our engine provides an additional speed boost due to its compact size, which allows for a much faster dose calculation speed than for a large size network. The student model decreases the inference time of the teacher model by 31.3%, 38.9%, and 60.0% on an A100 GPU, a RTX 3080 GPU, and a CPU, respectively. It should be noted that the reduction of computation time becomes more prominent when the DL-based dose calculation is performed on a CPU-only computing device. A better time-saving factor is therefore anticipated when low-powered computing devices are used for calculations.

Unlike some of the DL-based engines (Kontaxis *et al* 2020, Liu *et al* 2021, Tsekas *et al* 2021), which can only calculate an individual segment per inference, the proposed dose engine is able to provide a composite VMAT arc dose at one time using the composite doses in water as the input. Given that hundreds of sampled CPs are typically included in a VMAT arc, our dose engine demonstrates a great advantage over the segment/CP-based DL dose engines. It therefore leads to a substantial reduction in the overall calculation time. Moreover, adopting dose computation in a homogeneous medium (water) as the input allows for a faster preprocessing time than do DL engines requiring heterogeneity dose computation on CT for input generation.

During training, the same group of VMAT arcs were re-computed on different CT data sets to augment the dosimetric input dataset. Figure 11(a) demonstrates the dosimetric results of a VMAT arc calculated on its original planning patient anatomy, whereas the calculation results for another patient's anatomy is shown in

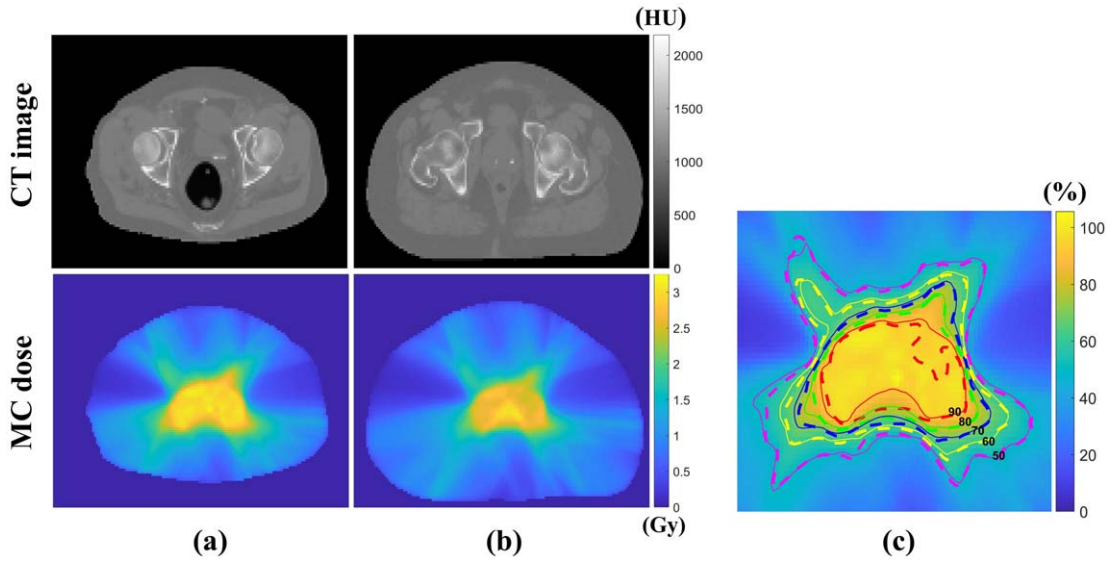

Figure 11. The identical VMAT arc calculated on (a) its original planning patient anatomy and (b) another patient's anatomy. (c) The 50% to 90% isodose comparisons between the doses of (a) solid lines and of (b) dashed lines. The doses were normalized to the treatment isocenter of the corresponding patient body. Dose comparison map was cropped and only included the isodose lines within the patient body in (a).

figure 11(b). Noticeable difference of isodose distribution between the two scenarios was observed, as shown in figure 11(c). These dosimetric differences are due to the patient anatomy and position variations. Therefore, the dose information obtained from same arcs but calculated on different CT data sets can be treated as additional dosimetric inputs to effectively increase the pool of the training data. This recycling strategy is able to alleviate the burden of massive data collection required by traditional DL frameworks. On the other hand, when the pool of accessible raw data are too small for adequate training, the recycling strategy can also be an alternative solution to expand the pool of training data. From a training efficacy point of view, the recycling strategy helps the DL model to better 'understand' the correlation between dose deposition and anatomical variance. It can therefore improve the robustness of the model.

The implementation of the KD framework in this study helps improve the performance of a small-capacity DL-based dose engine. This framework involves one extra DL-based calculation of a large-capacity model and one extra distillation loss calculation, as compared to a regular training framework. This results in a longer model training time. For example, the training times per epoch (about 88 iterations with a training batch size of 16) on a single A100 GPU were about 55.8 s, 51.3 s, and 38.3 s for the teacher, student, and student-only models, respectively. Accordingly, a prolonged training session can be expected when the KD framework is implemented.

From this proof-of-principle study, we found that the proposed DL-based dose calculation framework effectively works for prostate cancer cases with promising results. The feasibility of extending the proposed framework to other treatment sites (e.g. head and neck, lung, etc) and implementing other model compression techniques (e.g. parameters pruning and quantization) will be empirically investigated in a future study.

4.3. Potential clinical applications of the proposed dose engine

The repetitive dose calculations are generally required by iterative optimization processes of VMAT planning. The DL-based dose engines with their ultra-fast calculation speed allow a considerable reduction in the planning time. One research group (Liu *et al* 2021) has successfully integrated their DL-based dose engine (large-scale convolutional neural network (CNN)) into direct aperture optimization for IMRT inverse planning. They found that using the DL-based engine to calculate dose for each iteration can substantially reduce the planning time (up to 53% on average) and minimize the discrepancies between the optimized doses and the final plan doses (computed by high-accuracy MC algorithm). Since our method demonstrates the capability of decreasing inference time (39% and 60% on a GPU and a CPU, respectively) from a large-scale CNN engine calculation, a further reduction in the planning time is conceivable with the proposed KD-based DL engine. An engine with a fast computation speed is particularly desirable for online adaptive treatment planning. It can also be utilized as a fast and independent dose calculator for patient-specific quality assurance (QA) and TPS QA.

5. Conclusion

In this work, an ultra-fast DL-based dose engine for prostate VMAT plans was developed. The KD framework was implemented to improve the performance of the small-capacity dose engine. The results of dosimetric and inference time comparisons suggest that the proposed dose engine can perform highly-accurate VMAT dose calculations with a much faster calculation speed than traditional dose engines. Its small-capacity size further alleviates the computational resource requirement for the DL-based dose calculation. Thus, the proposed dose engine seems well suited to clinical settings where only limited computational devices are commonly available. In addition, the dose engine can be implemented for treatment planning to achieve a superior-quality plan in a more efficient manner.

Acknowledgments

This research is partially supported by NSF Grant CMMI-2016571.

Conflict of interest

The authors declare no conflicts of interest for this publication.

ORCID iDs

Yu Yang https://orcid.org/0000-0002-0502-7603

References

Ahnesjö A 1989 Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media *Med. Phys.* **16** 577–92
Bai T, Wang B, Nguyen D and Jiang S 2021 Deep dose plugin: towards real-time Monte Carlo dose calculation through a deep learning-based denoising algorithm *Mach. Learn.: Sci. Technol.* **2** 025033
Bedford J L 2009 Treatment planning for volumetric modulated arc therapy *Med. Phys.* **36** 5128–38
Chen H, Craft D L and Gierga D P 2014a Multicriteria optimization informed VMAT planning *Med. Dosim.* **39** 64–73
Chen W Z, Xiao Y and Li J 2014b Impact of dose calculation algorithm on radiation therapy *World J. Radiol.* **6** 874–80
Fu J, Bai J, Liu Y and Ni C 2020 fast monte carlo dose calculation based on deep learning 13th international congress on image and signal processing *BioMed. Eng. Inform. (CISP-BMEI)* **2020** 721–6
Hinton G E, Vinyals O and Dean J 2015 Distilling the knowledge in a neural network arXiv:abs/1503.02531
Kawrakow I 2000 Accurate condensed history monte carlo simulation of electron transport: I. EGSnrc, the new EGS4 version *Med. Phys.* **27** 485–98
Kontaxis C, Bol G H, Lagendijk J J W and Raaymakers B W 2020 Deep dose: towards a fast dose calculation engine for radiation therapy using deep learning *Phys. Med. Biol.* **65** 075013
Li Y, Rodrigues A, Li T, Yuan L, Yin F F and Wu Q J 2015 Impact of dose calculation accuracy during optimization on lung IMRT plan quality *J. Appl. Clin. Med. Phys.* **16** 219–28
Liu C, Ni X, Jin X and Si W 2021 NeuralDAO: incorporating neural network generated dose into direct aperture optimization for end-to-end IMRT planning *Med. Phys.* **48** 5624–38
Lu W and Chen M 2010 Fluence-convolution broad-beam (FCBB) dose calculation *Phys. Med. Biol.* **55** 7211–29
Peng Z, Shan H, Liu T, Pei X, Zhou J, Wang G and Xu X G 2019 Deep learning for accelerating Monte Carlo radiation transport simulation in intensity-modulated radiation therapy *Med. Phys.* 1–15 arXiv:1910.07735
Shepard D M, Earl M A, Li X A, Naqvi S and Yu C 2002 Direct aperture optimization: a turnkey solution for step-and-shoot IMRT *Med. Phys.* **29** 1007–18
Tsekas G, Bol G H, Raaymakers B W and Kontaxis C 2021 Deep dose: a robust deep learning-based dose engine for abdominal tumours in a 1.5 T MRI radiotherapy system *Phys. Med. Biol.* **66** 065017
Unkelbach J *et al* 2015 Optimization approaches to volumetric modulated arc therapy planning *Med. Phys.* **42** 1367–77
Xing Y, Nguyen D, Lu W, Yang M and Jiang S 2020a Technical note: a feasibility study on deep learning-based radiotherapy dose calculation *Med. Phys.* **47** 753–8
Xing Y, Zhang Y, Nguyen D, Lin M H, Lu W and jiang S 2020b Boosting radiotherapy dose calculation accuracy with deep learning *J. Appl. Clin. Med. Phys.* **21** 149–59
Zhang J, Liu S, Yan H, Li T, Mao R and Liu J 2020 Predicting voxel-level dose distributions for esophageal radiotherapy using densely connected network with dilated convolutions *Phys. Med. Biol.* **65** 205013