Technical Note: Dose prediction for radiation therapy using feature-based losses and One Cycle Learning

Lukas Zimmermann
Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria

Erik Faustmann and Christian Ramsl
Technical University of Vienna, Vienna, Austria

Dietmar Georg and Gerd Heilemann[a)]
Department of Radiation Oncology, Medical University of Vienna, Vienna, Austria

(Received 22 November 2020; revised 14 January 2021; accepted for publication 29 January 2021; published 22 June 2021)

Purpose: To present the technical details of the runner-up model in the open knowledge-based planning (OpenKBP) challenge for the dose–volume histogram (DVH) stream. The model was designed to ensure simple and reproducible training, without the necessity of costly advanced generative adversarial network (GAN) techniques.
Methods: The model was developed based on the OpenKBP challenge dataset, consisting of 200 and 40 head-and-neck patients for training and validation, respectively. The final model is a U-Net with additional ResNet blocks between up- and down convolutions. The results were obtained by training the model with AdamW with the One Cycle scheduler. The loss function is a combination of the L1 loss with a feature loss, which uses a pretrained video classifier as a feature extractor. The performance was evaluated on another 100 patients in the OpenKBP test dataset. The DVH metrics of the test data were evaluated, where $D_{0.1cc}$, and D_{mean} were calculated for the organs at risk (OARs) and $D_{1\%}$, $D_{95\%}$, and $D_{99\%}$ were computed for the target structures. DVH metric differences between predicted and true dose are reported in percentage.
Results: The model achieved 2nd and 4th place in the DVH and dose stream of the OpenKBP challenge, respectively. The dose and DVH score were 2.62 $\pm$ 1.10 and 1.52 $\pm$ 1.06, respectively. Mean dose differences for the different structures and DVH parameters were within $\pm 1\%$.
Conclusion: This straightforward approach produced excellent results. It incorporated One Cycle Learning, ResNet, and feature-based losses, which are common computer vision techniques. [https://doi.org/10.1002/mp.14774]

Key words: deep learning, dose prediction, radiation therapy

1. INTRODUCTION

Treatment techniques in radiation therapy have become more complex and time consuming. As a result, there is demand for automating the planning process to provide fast and reliable treatment plans with consistent plan quality. A promising method to help creating these plans is deep learning, which has been used to predict dose distributions for various body sites.[1–3] Unfortunately, most models are developed and tested on private datasets, which makes it difficult to compare the methods. The open knowledge-based planning (OpenKBP) challenge provided a platform with a dataset of intensity modulated radiation therapy treatment plans to make a quantitative comparison of these models possible.[4] This technical note provides detailed information on the first runner-up model in the dose–volume histogram (DVH) stream of the OpenKBP challenge. Recent dose prediction models have utilized generative adversarial networks (GANs) to improve conversion performance.[1,3] However, despite the fact that GANs resulted in great success for many different applications, the technique still suffers from shortcomings like vanishing gradients and mode collapse. This makes GANs difficult to train and requires an extensive hyperparameter search for optimal performance. Because of this we focused our work on alternative techniques which include learning rate scheduling,[5] advanced nonlinear activation functions,[6] and neural network based feature cost functions[7] to make predictions more accurate and reproducible during testing.

The baseline U-Net model has been used for dose prediction within a GAN architecture that was based on pix2pix.[1,8] Even though various GAN techniques (e.g., conditional GANs,[1,8] spectral normalization,[9] attention-gated convolutions[10]) help to improve performance, the hyperparameter search is difficult as there are many hyperparameters. Multiple research groups demonstrated that a simple model like the U-Net or pretrained ResNet classifiers can be trained without the requirement of GAN training and still perform well in the head and neck region.[2,11–15] Most of the studies utilized the mean squared error loss function for optimization, which does not capture geometric representations of the volume. DVH parameters for these studies ranged from 3 to 5% of the mean or median value relative to the prescribed

dose, which is similar to values reported for commercially available systems.[16,17] However, the results of published studies are difficult to compare due to different datasets and non-standardized evaluation procedures.

In this study, the focus was to use robust training techniques which do not rely on GAN methods. For example, fast.ai is a library that uses One Cycle Learning Scheduling which is a method that overcomes warm-up problems of the Adam optimizer, to efficiently train models without GAN techniques.[5,18]

Additionally, optimizers have been studied extensively in the past years which has resulted in a large variety of methods. AdamW, for example, combines the Adam optimizer with decoupled weight decay regularization to prevent overfitting.[19]

Another important factor for neural networks is the choice of the activation function. Many new activation functions have been developed in an effort to replace the popular ReLU activation function, which is non-differentiable at zero and results in no gradient flow if the input value is below zero. One of these new activation functions is Mish, which does not saturate and is everywhere differentiable and demonstrated improved accuracy by 1–2% for image classification and object detection.[6]

Lastly, many new loss functions have been developed that incorporate feature loss metrics, which are extracted via pre-trained models.[7] Feature losses were already successfully applied for denoising images like low- dose CT, tomography, and optical coherence tomography.[20–23] A previous study of Ngyuen et al. demonstrated good results when a DVH based loss was included for dose prediction in the pelvic region which highlights the necessity of domain specific loss functions.[24]

2. MATERIALS AND METHODS

2.A. Datasets

The OpenKBP provided a preprocessed dataset with 200, 40, and 100 patients for training, validation, and testing, respectively. All patients included a CT, structure sets for organs at risk (OAR) and target volumes (with up to three dose levels), as well as corresponding dose distributions. All volumes are represented as $128 \times 128 \times 128$ matrix with a variable voxel resolution of about $3.5 \times 3.5 \times 2\,mm^3$.

2.B. Baseline model

The baseline model was the U-Net implementation of the pix2pix model.[8] However, three major changes were made to that baseline model: (a) three-dimensional (3D) convolutions were used instead of two-dimensional convolutions, (b) instance normalization with affine transformations instead of batch normalization, and (c) a sigmoid output function instead of a tanh.

The model input consisted of the challenge data as a concatenated 4D volume. CT and dose distribution values were clipped and divided by 4000 and 100, respectively, to bring the voxel values into the interval between [0, 1]. An illustration of the final model can be found in Fig. 1 and the code for the model is available at https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.

2.C. Hyperparameter search

All different configurations of the parameters were tracked using the weights and biases service (Weights&Biases Software wandb.com, CA, USA) to allow for fast and easy evaluation. Starting from the baseline model, the following hyperparameters were adapted:

- **Activation function:** Set to Mish for all convolutions, except for the output of the network.
- **ResNet blocks:** Between the up- and down-sampling blocks a ResNet block was included with a bottleneck. Feature size was reduced to 64 with a 1×1 convolution, followed by a 3×3 convolution with 64 features, and a last 1×1 convolution increasing the feature size to the initial one. The convolution blocks follow the form of convolution, normalization, and activation where the activation is skipped for the final convolution.
- **Masking:** To increase the information density, the error calculation was limited to the external contour.
- **Loss function:** For training, the L1 metric was used as a baseline. Furthermore, a feature based loss was implemented which extracted high higher-order information with a pre-trained model. Since the data is three dimensional, the pre-trained ResNet3D for video classification was utilized for feature extraction.[25] This pre-trained model was taken from torchvision.[26] To fit the input criteria of the classifier, the predicted dose output and the ground truth dose were repeated three times to fill the red-green-blue (RGB) channels. Features were extracted from different depths of the model, where the single block outputs were used which are called, for example, *stem*, *layer1*, *layer2* as can be seen in Table I. For more information, please see the original paper[25] or the torchvision implementation.[26] The model outputs can be seen as dimensionality reduction of the input dose matrix into a lower dimension matrix. By passing the ground truth and the prediction through the model the output features can be compared by a distance measure (see Fig. 1). The overall loss function is then given as:

$$L = \lambda \frac{1}{m} \|x - y\|_1 + \sum_{i=1}^{n} \frac{1}{m} \lambda_{F,i} \|F^{[i]}(x) - F^{[i]}(y)\|_1 \quad (1)$$

where $F(\cdot)$ is the pre-trained model, m is the number of voxels considered in the comparison, $[i]$ denotes the last layer positions of the the different blocks specified above, and x,y are the ground truth and predicted dose distributions, respectively. λ and $\lambda_{F,i}$ are weighting factors where λ was set to 100 for all experiments and $\lambda_{F,i}$ to 1.

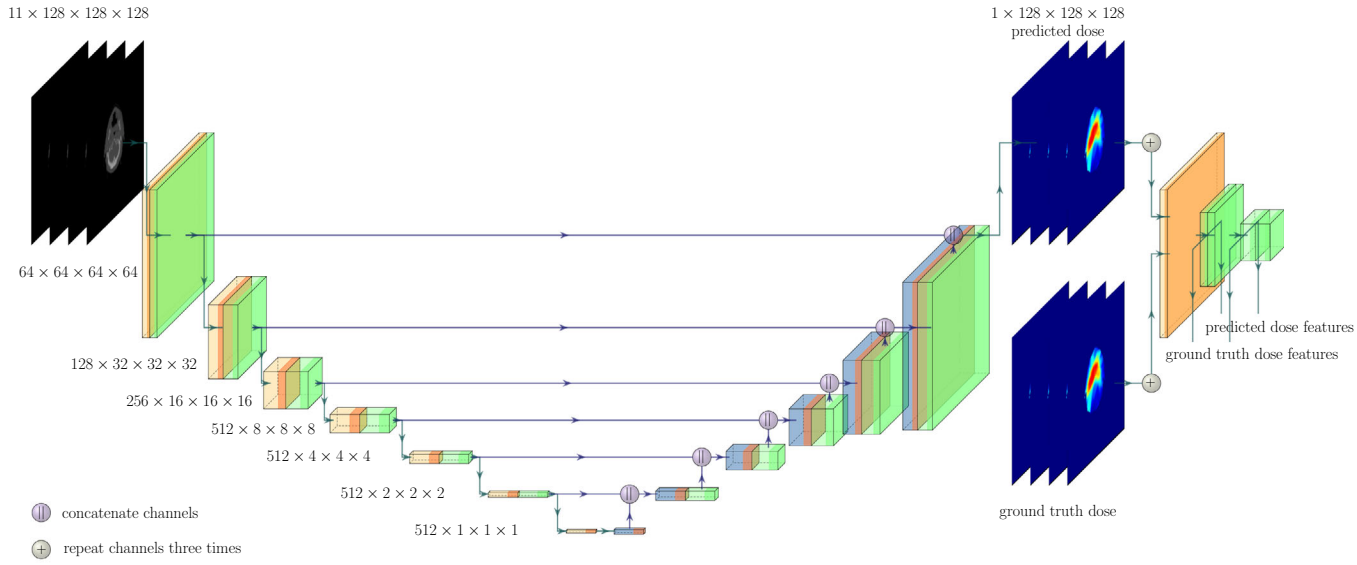

FIG. 1. The final U-Net model with ResNet blocks (green), 4 × 4 convolutions (orange) and transposed convolutions (blue). Purple dots represent the feature map concatenation. Note that the blocks represent the output of the respective operations given by the color. On the right side of the predicted dose, the pre-trained ResNet model can be seen where the feature outputs are extracted from different depths for the predicted and the ground truth dose distribution. Please note that the ResNet blocks of the U-Net and the pre-trained classifier differ in architecture. [Color figure can be viewed at wileyonlinelibrary.com]

2.D. Model training

The maximum learning rate was defined with the learning rate finder which resulted in 10^{-3} for all different configurations. The models were trained using AdamW with a weight decay of 10^{-4} and β_1 and β_2 values of 0.5 and 0.999, respectively. The One Cycle Learning rate schedule was applied over a training time of 200 epochs. Additionally, data augmentation was performed including transversal flips with orientation consistency, and random translations of the image volume along all three coordinate axes. The algorithm was implemented using Pytorch Lightning with Pytorch.

2.E. Evaluation

The model performance was tracked with the dose score which is the mean absolute error of the predicted dose distribution inside of the external contour as well as the DVH score which is the mean absolute error of five predefined DVH metrics ($D_{0.1cc}$, D_{mean}, $D_{1\%}$, $D_{95\%}$, and $D_{99\%}$). The mean and standard deviation of the dose and DVH score over all patients are reported. Improvements are given via the difference to the means of the baseline model for the dose and DVH score. The best performing model was chosen by the lowest dose and DVH score of the validation dataset.

TABLE I. Convolution operations in the ResNet3D. All layers use batch normalization and ReLU. The encoder number for F gives the integer output position of the formula.

Encoder number F	Layer name	Output size	Convolution layers
F^0	*stem*	64 × 128 × 64 × 64	3 × 7 × 7, 64, stride 1 × 2 × 2
F^1	*layer1*	64 × 128 × 64 × 64	$\begin{bmatrix} 3\times3\times3, 64 \\ 3\times3\times3, 64 \end{bmatrix}\times2$
F^2	*layer2*	128 × 64 × 32 × 32	$\begin{bmatrix} 3\times3\times3, 128 \\ 3\times3\times3, 128 \end{bmatrix}\times2$

The dose and DVH score were computed and reported for the test dataset for the best performing model. All relative DVH metric differences are additionally reported for the test dataset, where $D_{0.1cc}$ and D_{mean} are computed for the OARs and $D_{1\%}$, $D_{95\%}$ and $D_{99\%}$ are computed for the target structures. The dose differences for the single structures are given as:

$$\Delta D = \frac{D_{pred} - D_{true}}{70Gy} \times 100 \quad (2)$$

where D_{pred} and D_{true} are the predicted and ground truth dose metric, respectively.

3. RESULTS

The baseline model yielded relatively good starting scores for both dose and DVH, which were further improved by the proposed methods (see Table II). The final model that performed best used ResNet blocks, Mish activation functions, mask guided loss metrics, and the first convolution block (*stem*) of the feature loss.

This configuration resulted in a dose score of 2.62 ± 1.10 (*4th* place in dose stream) and a DVH score of 1.52 ± 1.06. (2nd place in DVH stream) on the test set of 100 patients. By omitting one patient of the test dataset, which was identified to include a wrong PTV label, the final dose score and the DVH score could be improved to 2.58 ± 1.04 and 1.43 ± 0.53, respectively.

TABLE II. Dose and dose-volume histogram (DVH) scores for the different hyperparameter settings and their mean differences to the baseline model (Δ) for the validation dataset. The highest difference is shown in bold.

	Dose score	Δ	DVH score	Δ
Baseline	2.651 ± 0.849	–	1.666 ± 0.853	–
+ ResNet blocks	2.548 ± 0.796	0.103	1.617 ± 0.759	0.049
+ Mish	2.534 ± 0.796	0.117	1.611 ± 0.777	0.055
+ Masking	2.530 ± 0.747	0.121	1.607 ± 0.789	0.059
+ Feature loss	2.503 ± 0.738	**0.148**	1.563 ± 0.790	**0.103**

Figure 2 shows the percentage dose difference between the predicted and ground truth DVH metric. The medians of all parameters were distributed between −1.2% and 0.9%. The inter-quartile ranges were also within −4% and 3%. In total, there were 61 outliers across all DVH metrics, however, only 16% outliers were observed that had a dose difference > ± 10%.

4. DISCUSSION AND CONCLUSION

The initial U-Net baseline model trained with AdamW and One Cycle Learning resulted already in good results. We observed that training the model for 100 epochs is efficient as the validation loss improved only slightly by increasing the epochs to 200. Two of the main effects of One Cycle Learning are the adaptive changing of the learning rate and the momentum, which were recently shown to be a key component for improving optimization.[27] Additionally, instance normalization with affine transformations was applied because initial experiments with batch normalization showed worse results during validation due to memory dependent small batch size (n = 1).

The modifications to the baseline model affected the dose and DVH score differently. The use of ResNet blocks and masking resulted in much better dose score (improvement of 0.103 and 0.121, respectively). However, the feature loss helped to improve the DVH score by 0.103 and consequently the DVH metrics further, even though the dose score changed only slightly. The model's feature extraction network learned from non-medical RGB video data, which are not representative for dose distributions. Still the model seems to provide useful information to improve model performance. Using the first layer of the pre-trained model resulted in the best performance and including one additional layer output (*layer1*) resulted in a similar dose score but a worse DVH score value. Models trained with this feature loss technique would likely benefit if the corresponding classifier was pre-trained on medical images specific to the task at hand. This should be further investigated as great success was achieved by using feature- based losses not only in this study but also in other fields of application.

There are many other promising techniques from the field of computer vision that could be explored. For example, transfer learning with 3D models was successfully used for classification of medical data.[28] Similarly, the model could be adapted to be used as a pre-trained encoder part of a U-Net, as it is performed by the fast.ai library.[18] Unfortunately, these kind of experiments were excluded due to memory and time constrains.

Changing the activation function also slightly improved the performance as can be seen in Table II. Since the Mish implementation was done in native pytorch, memory

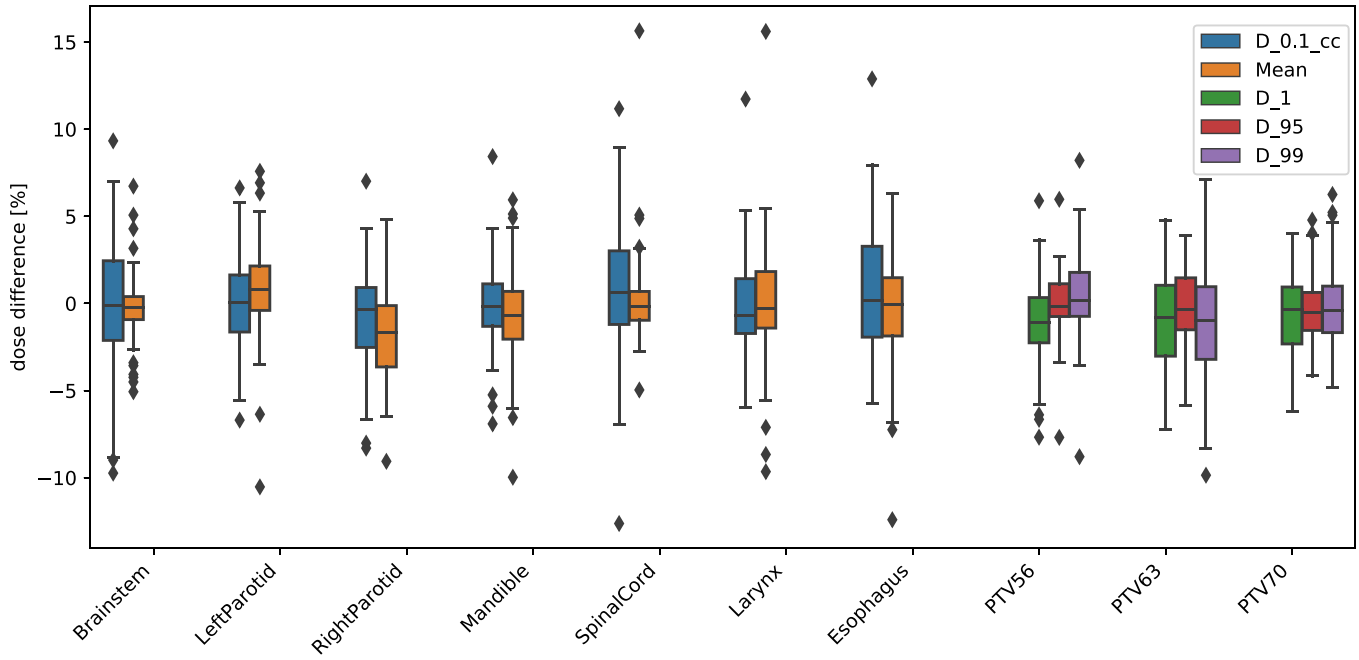

FIG. 2. The colored dose parameters for the test dataset for all organs at risks and target structures. The boxes indicate median and interquartile range (IQR) and the whiskers extend to 1.5 times the IQR. Outliers are denoted by diamonds. [Color figure can be viewed at wileyonlinelibrary.com]

requirements were sub-optimal and high compared to ReLU based activation functions. The entire implementation used to obtain the training results in wandb as well as the link to the wandb report can be found at this github repository.

In conclusion, this straightforward approach resulted in good results for dose predictions, with only common computer vision techniques and without the necessity of complicated training methods (e.g., GANs).

[a)]Author to whom correspondence should be addressed. Electronic Email: gerd.heilemann@meduniwien.ac.at.

REFERENCES

1. Babier A, Mahmood R, McNiven AL, Diamant A, Chan TCY. Knowledge-based automated planning with three-dimensional generative adversarial networks. *Med Phys*. 2020;47:297–306.
2. Nguyen D, Long T, Jia X, et al. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci Rep*. 2019;9:1–10.
3. Kearney V, Chan JW, Wang T, et al. DoseGAN: a generative adversarial network for synthetic dose prediction using attention-gated discrimination and generation. *Sci Rep*. 2020;10:1–8.
4. Babier A, Zhang B, Mahmood R, et al. OpenKBP: The open-access knowledge-based planning grand challenge. arXiv. 2020.
5. Leslie NS, Nicholay T. Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates. arxiv; 2017.
6. Misra D. Mish: A Self Regularized Non-Monotonic Activation Function. BMVC, 2020.
7. Johnson J, Alahi A, Fei-Fei LI. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. ECCV. 2016.
8. Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. Proceedings - 30th IEEE Conference on Computer Vision and Pattern RecognitionCVPR, 2017, 2017:5967–5976.
9. Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. ICLR, 2018.
10. Brock A, Donahue J, Simonyan K. Large -Scale Tracing of GAN training for high fidelity natural image synthesis. ICLR. 2019.
11. Dan N, Xun J, David S, et al. Three-dimensional radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected u-net deep learning architecture. *Phys Med Biol* 2019:64.
12. Liu Z, Chen X, Men K, Yi J, Dai J. A deep learning model to predict dose-volume histograms of organs at risk in radiotherapy treatment plans. *Med Phys*. 2020;47:5467–5481.
13. Liu Z, Fan J, Li M, et al. A deep learning method for prediction of three-dimensional dose distribution of helical tomotherapy. *Med Phys*. 2019;46:1972–1983.
14. Fan J, Wang J, Chen Z, Chaosu HU, Zhang Z, Weigang HU. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Med Phys*. 2019;46:370–381.
15. Chen X, Men K, Li Y, Yi J, Dai J. A feasibility study on an automated method to generate patient-specific dose distributions for radiotherapy using deep learning. *Med Phys*. 2019;46:56–64.
16. McIntosh C, Welch M, McNiven A, Jaffray DA, Purdie TG. Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method. *Phys Med Biol*. 2017;62:5926–5944.
17. Cornell M, Kaderka R, Hild SJ, et al. Noninferiority study of automated knowledge-based planning versus human-driven optimization across multiple disease sites. *Int J Radiat Oncol Biol Phys*. 2020;106(2):430–439.
18. Howard J, Gugger S. Fastai: a layered API for deep learning. *Information*. 2020;11:1–26.
19. Loshchilov I, Hutter F. Decoupled weight decay regularization. ICLR, 2019. https://arxiv.org/abs/1711.05101v3
20. Yang Q, Yan P, Zhang Y, et al. Low-dose ct image denoising using a generative adversarial network with Wasserstein distance and perceptual loss. *IEEE Trans Med Imaging*. 2018;37:1348–1357.
21. Li M, Hsu W, Xie X, Cong J, Gao W. Sacnn: Self-attention convolutional neural network for low-dose ct denoising with self-supervised perceptual loss network. *IEEE Trans Med Imaging*. 2020;39:2289–2301.
22. Qiu B, Huang Z, Liu XI, et al. Noise reduction in optical coherence tomography images using a deep neural network with perceptually-sensitive loss function. *Biomed Opt Express*. 2020;11:817–830.
23. Gao M, Samala RK, Fessler JA, Chan H-P. Deep convolutional neural network denoising for digital breast tomosynthesis reconstruction. In: Chen G-H, Bosmans H, eds. *Medical Imaging 2020: Physics of Medical Imaging*. SPIE: International Society for Optics and Photonics; 2020;11312:173–178.
24. Nguyen D, McBeth R, Barkousaraie AS, et al. Incorporating human and learned domain knowledge into training deep neural networks: a differentiable dose-volume histogram and adversarial in-spired framework for generating Pareto optimal dose distributions in radiation therapy. *Med Phys*. 2020;47:837–849.
25. Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M. A closer look at spatiotemporal convolutions for action recognition. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018:6450–6459.
26. Sébastien M, Yann R. Torchvision the machine-vision package of torch. In Proceedings of the 18th ACM International Conference on MultimediaMM '10, Association for Computing Machinery: New York, NY, USA. 2010:1485-1488.
27. Defazio A. Understanding the role of momentum in non-convex optimization: Practical insights from a Lyapunov analysis, 2020.
28. Rajpurkar P, Park A, Irvin J, et al. AppendiXNet: Deep learning for diagnosis of appendicitis from a small dataset of CT exams using video pretraining. *Sci Rep*. 2020;10:2045–2322.