HHS Public Access

Author manuscript

Med Phys. Author manuscript; available in PMC 2021 June 01.

Published in final edited form as:

Med Phys. 2020 June ; 47(6): 2329–2336. doi:10.1002/mp.14114.

Operating a Treatment Planning System using a Deep-Reinforcement-Learning based Virtual Treatment Planner for Prostate Cancer Intensity-Modulated Radiation Therapy Treatment Planning

Chenyang Shen[1,2], **Dan Nguyen**[1], **Liyuan Chen**[1], **Yesenia Gonzalez**[1,2], **Rafe McBeth**[1], **Nan Qin**[1,2], **Steve B. Jiang**[1], **Xun Jia**[1,2]

[1]Medical Artificial Intelligence and Automation (MAIA) Laboratory, Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

[2]innovative Technology Of Radiotherapy Computation and Hardware (iTORCH) Laboratory, Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

Abstract

Purpose: In the treatment planning process of intensity modulated radiation therapy (IMRT), a human planner operates the treatment planning system (TPS) to adjust treatment planning parameters, e.g. dose volume histogram (DVH) constraints' locations and weights, to achieve a satisfactory plan for each patient. This process is usually time-consuming, and the plan quality depends on planer's experience and available planning time. In this study, we proposed to model the behaviors of human planners in treatment planning by a deep reinforcement learning (DRL)-based virtual treatment planner network (VTPN), such that it can operate the TPS in a human-like manner for treatment planning.

Methods and Materials: Using prostate cancer IMRT as an example, we established the VTPN using a deep neural network developed. We considered an in-house optimization engine with a weighted quadratic objective function. VTPN was designed to observe an intermediate plan DVHs and decide the action to improve the plan by changing weights and threshold dose in the objective function. We trained the VTPN in an end-to-end DRL process in 10 patient cases. A plan score was used to measure plan quality. We demonstrated the feasibility and effectiveness of the trained VTPN in another 64 patient cases.

Results: VTPN was trained to spontaneously learn how to adjust treatment planning parameters to generate high-quality treatment plans. In the 64 testing cases, with initialized parameters, quality score was 4.97 (±2.02), with 9.0 being the highest possible score. Using VTPN to perform treatment planning improved quality score to 8.44 (±0.48).

Conclusions: To our knowledge, this was the first time that intelligent treatment planning behaviors of human planner in external beam IMRT are autonomously encoded in an artificial

Corresponding Author: Xun Jia, PhD (Xun.Jia@UTSouthwestern.edu), Chenyang Shen, PhD (Chenyang.Shen@UTSouthwestern.edu).

intelligence system. The trained VTPN is capable of behaving in a human-like way to produce high-quality plans.

1. Introduction

Inverse treatment planning of a modern radiation therapy technique, e.g. Intensity Modulated Radiation Therapy (IMRT)[1], is one of the most critical steps[2, 3]. The optimization problem of inverse planning typically contains multiple terms and constraints designed for various clinical or practical considerations, as well as a set of treatment planning parameters (TPPs) such as weights and locations of dose-volume histogram constraints. Although modern treatment planning systems can solve the optimization problem for given TPPs, deciding these TPP values is typically beyond their capability. Nonetheless, TPPs are patient-specific and of critical importance for the creation of a high-quality plan. To date, human-based approaches are still the main solution to this problem: planners operate treatment planning systems to create satisfactory plans by setting and adjusting TPPs based on human intuitions. This is not only tedious and time-consuming, the final plan quality is affected by many factors, such as the experience of the planner and the available time. Plans are often unwittingly accepted under the time pressure, although further improvement might be still possible. Therefore, an automatic planning process that is able to determine the TPP values to achieve high-quality plans is highly desired.

Extensive research efforts have been devoted to solving this problem. One of the most commonly used techniques was to add a loop of TPP optimization on top of the optimization algorithm performed with a fixed set of TPP values[4–7]. Heuristic approaches have been employed to handle the TPP optimization problem, such as assigning values based on the location information of voxels[8–10]. Fuzzy inference techniques were proposed to adjust the TPPs[11, 12]. Recently, statistical methods have also been introduced to establish the relationship between TPPs and patient anatomy[13] by analyzing previous delivered high quality plans[14, 15].

A human planner usually takes a trial-and-error approach to adjust TPPs. By observing a plan produced by the planning system under a set of TPPs, the planner decides how to change them to improve plan quality. This process continues to gradually achieve a satisfactory plan. Adjusting the TPPs can be viewed as a decision-making problem. Recently, tremendous success in artificial intelligence has been achieved across different disciplines[16–27] including radiation oncology[28–30]. More importantly, pioneer work showed that a system built under the deep reinforcement learning (DRL) framework is able to make decisions for certain tasks in a human-like fashion, such as playing Atari games or Go, achieving performance similar to, or even better than humans[21, 26, 27]. In light of these successes, we have proposed a novel treatment planning framework called *intelligent automatic treatment planning*. More specifically, a virtual treatment planner built under DRL can be trained to learn how to operate an optimization engine to generate high quality plans. We have demonstrated the feasibility and potential of this approach in a proof-of-principle study of inverse treatment planning for high dose-rate brachytherapy[30]. The study was limited to a small-scale optimization problem in brachytherapy and the DRL system was designed to only adjust organ weights in the relatively simple objective function.

In this paper, we further explore the feasibility of developing the virtual treatment planner that can adjust more general TPPs for external beam radiotherapy. We focus on prostate cancer IMRT, as this problem involves a relatively small number of critical organs. To our knowledge, this is the first time that a computational tool is built to have intelligence in external beam radiotherapy treatment planning regime, and is trained to be capable of operating an optimization engine in a human-like fashion to create high quality plans.

2. Methods and Materials

2. 1 Overview of the intelligent automatic treatment planning framework

In a typical inverse treatment planning workflow, a planner repeatedly observes plans created by the optimization engine and adjusts TPPs to gradually generate a satisfactory plan. We propose to develop a deep neural network named virtual treatment planner network (VTPN), trained to learn the parameter-adjustment policy via an end-to-end DRL process. The trained VTPN decides the way of changing TPPs to improve plan quality in lieu of a human planner in the planning process, as shown in Fig. 1.

2.2 Treatment planning optimization algorithm

In this paper, we consider the following in-house developed engine targeting at the fluence map optimization problem[31]:

$$\begin{aligned} &\min_{x} \frac{1}{2}\|Mx-d_p\|_-^2+\frac{\lambda}{2}\|Mx-d_p\|_+^2+\sum_i \frac{\lambda_i}{2}\|M_i x-\tau_i d_p\|_+^2, \\ &\text{s.t. } x\geq 0, D_{95\%}(Mx)=d_p. \end{aligned} \tag{1}$$

$\|\cdot\|_-$ and $\|\cdot\|_+$ indicate l_2 norm computed only for negative and positive elements, respectively. $x \geq 0$ is the beam fluence map to be determined. M and M_i are dose deposition matrices for planning target volume (PTV) and the i-th organ at risk (OAR). d_p denotes prescription dose. λ and λ_i are weighting factors to penalize overdose to PTV and i-th OAR, respectively. τ_i adjusts threshold dose of the i-th OAR. The hard constraint $D_{95\%}(Mx) = d_p$ indicates that 95% of PTV is required to receive dose no lower than the prescription dose. In this model, λ, λ_i, and τ_i^d are TPPs to be adjusted to yield a satisfactory plan. For the example problem of prostate cancer IMRT, we consider only bladder and rectum as OARs. Hence, there are five TPPs, i.e. λ, $\lambda_{bladder}$, λ_{rectum}, $\tau_{bladder}$, and τ_{rectum}. With a given set of TPPs, this optimization problem is solved using alternating direction method of multipliers (ADMM)[32, 33].

2.3 Virtual treatment planner network

VTPN observes a treatment plan generated by the optimization engine under a given set of TPPs, and outputs the TPP to adjust, as well as the direction and magnitude of adjustment. We employ the Q-learning framework[34], which considers the following optimal action-value function:

$$Q^*(s,a)=\max_{\pi}[r^l+\gamma r^{l+1}+\gamma^2 r^{l+2}+\cdots|s^l=s,a^l=a,\pi]. \tag{2}$$

Q^* is a function of system state s, i.e. the observed treatment plan, and action a, i.e. the decision regarding which TPP to adjust and how to adjust it. s^l and a^l stand for the state and action at the l-th TPP adjustment step, respectively. r^l is the reward obtained at step l, which is given by a pre-defined reward function related to clinical objectives. A positive reward is obtained, if the clinical objectives are better met by applying the action a^l on the state s^l, namely if a better plan is obtained with the updated TPP, and negative otherwise. $\gamma \in [0, 1]$ is a discount factor used to emphasize more on the current reward as opposed to future reward. $\pi = P(a|s)$ denotes the policy of TPP adjustment: taking an action a based on the observed state s. The goal of developing the policy is to establish the Q^* function. Once it is determined, the policy can be decided as $a = \arg\max_{a'} Q^*(s, a')$, namely, choosing the action a to maximize the Q^* function for the observed plan state s.

The general form of the Q^* function is unknown and hence we employ the powerful deep neural network as a parametrization form[21, 22]. As such, we construct a deep neural network VTPN, denoted as $Q(s, a; W)$ with network parameters W to be determined via training. The network structure is shown in Fig. 2. More specifically, in total eight fully-connected layers and seven ReLU layers are incorporated to construct each subnetwork of VTPN. To define each fully-connected layer, a set of weights w_j^i and biases b_j^i are needed, where $i = 1,2,\ldots N$ is the index of subnetworks and $j = 1,2,\ldots,8$ gives the index of fully-connected layers. Then $W = \left\{w_1^1, b_1^1, w_2^1, b_2^1, \ldots, w_8^1, b_8^1, w_1^2, b_1^2, \ldots, w_8^N, b_8^N\right\}$ represents the union of the weights and biases from all the subnetworks. In this study, we choose the plan's dose volume histograms (DVH) as the input to the VTPN, as it is usually the starting point that a human planner uses to evaluate a plan quality. The input has three columns corresponding to the DVHs of PTV and two OARs. Considering that there are five TPPs to adjust, VTPN consists of five subnetworks, each designed for one TPP. For each one, we consider five possible TPP adjustment actions, i.e. increase or decrease its value by 50%, increase or decrease by 10%, and keep it unchanged. The ratios of change, i.e. 50% and 10% are arbitrary chosen, as we expect they would not critically affect parameter-tuning performance, but only the speed to reach convergence. Each subnetwork contains five output nodes, with their values being the Q^* function value for the corresponding action.

2.4 Training VTPN

2.4.1 Deep reinforcement learning—The training process via an end to end DRL process is based on Bellman equation[35], which is a general property of the optimal action-value function $Q^*(s, a)$:

$$Q^*(s, a) = r + \gamma \max_{a'} Q^*(s', a'). \quad (3)$$

s' is the state obtained by applying the action a to the current state s, while r is the reward obtained. In our case, s' stands for the new plan after solving the optimization problem using TPPs updated by the action a that is decided based on observing the current treatment plan s. As VTPN $Q(s, a; W)$ is an approximation of the Q^* function, we define a quadratic loss function with respect to W:

$$H(W) = [r + \gamma \max_{a'} Q(s', a'; W) - Q(s, a; W)]^2 . \quad (4)$$

Our goal is to determine W through DRL to minimize this loss function, which hence ensures Eq. (3) and therefore VTPN will approach the optimal action-value function Q^*. This is achieved by a two-loop learning strategy. For the inner loop, the following term is minimized with $\widehat{W}$ fixed:

$$L(W) = [r + \gamma \max_{a'} Q(s', a'; \widehat{W}) - Q(s, a; W)]^2 . \quad (5)$$

W is updated in a gradient descent form where the gradient of $L(W)$ can be simply derived as $\partial L(W)/\partial W = \left[Q(s, a; W) - r - \gamma \max_{a'} Q\left(s', a'; \widehat{W}\right)\right] \partial Q(s, a; W)/\partial W$. The last term $\partial Q(s, a; W)/\partial W$ can be computed via the standard back-propagation strategy[36]. Stochastic gradient descent[36] is implemented. In each step of the outer loop, $\widehat{W}$ is simply updated by setting it to the latest network parameters W. It is then fixed, and we move on to the optimization process of inner loop for the next iteration. Eventually, $\widehat{W}$ and W are expected to converge at the end of the learning process, making Eq. (4) minimized.

2.4.2. Reward function—By quantifying the plan quality using a scoring function ψ (higher is better), the reward function can be calculated as $r = \psi(s') - \psi(s)$. The reward explicitly measures the difference in plan quality between the two states before and after TPP adjustment. It is positive, if plan quality is improved, and negative otherwise. As we focus on the example problem of IMRT treatment planning for prostate cancer, we consider $\psi(s)$ as a scoring system of ProKnow (ProKnow Systems, Sanford, FL, USA) for prostate cancer IMRT plan. ProKnow scoring system used in this study is the equally-weighted sum over 9 selected scores for different clinical criteria including $D_{PTV}(0.3_{CC})$, $V_{bladder}(80Gy)$, $V_{bladder}(75Gy)$, $V_{bladder}(70Gy)$, $V_{bladder}(65Gy)$, $V_{rectum}(75Gy)$, $V_{rectum}(70Gy)$, $V_{rectum}(65Gy)$, and $V_{rectum}(60Gy)$. It does not include all the scoring criteria from the original ProKnow scoring system since we only consider those scores applicable to our study. For instance, since we set $D_{95\%}$ of PTV to be the hard constraint in the optimization problem, which is always strictly satisfied, we remove the score criterion corresponding to PTV $D_{95\%}$ in original ProKnow scoring system. For each criterion, the score is defined as a piecewise linear function, with maximum of 1 and minimum of 0. One example of scoring function regarding bladder sparing is shown in Fig. 3: one for $V_{bladder}$(80Gy) < 15%, zero for $V_{bladder}$(80Gy) > 20%, and linearly interpolated between 15 and 20%. Then the ProKnow score of a plan ranges between 0 and 9 in the scoring system used in this study, with higher score indicating better plan quality.

We further modify the scoring system for the training purpose to respect human's intentions. First, the step function form of ProKnow does not penalize further, once a criterion is satisfied to a certain extent. For the $V_{bladder}$(80Gy) criterion, the score does not differentiate plans with $V_{bladder}$(80Gy) < 15% or $V_{bladder}$(80Gy) > 20%. This contradicts a human planer's knowledge of reducing OAR dose as much as possible. To address this issue, we replace the step function by a smooth sigmoid-shape function shown in Fig 3, which preserves the general behavior of the step function but has a monotonically decaying form.

We make similar modifications to other criteria in ProKnow. Second, PTV dose is of top priority in treatment planning, which is evaluated by $D_{95\%}$ and D_{max} in ProKnow. Our optimization engine already treats $D_{95\%}$ as a hard constraint. For D_{max}, we empirically amplify the weight for its score by a factor of six to highlight its importance as compared to other criteria.

2.4.3 Training strategy—The training process is performed using 10 training patient cases in 100 episodes. In each episode, we start with all TPPs set to be unity for all training cases. For each training case, sequence of TPPs adjustment steps are then performed, each followed by a plan optimization process. We terminate the TPPs adjustment and move on to the next patient if the number of adjustment steps reaches the maximum of 25. In each step, we select an action to adjust a TPP using the ϵ-greedy algorithm. Specifically, with a probability of ϵ (initial $\epsilon = 0.99$), we select a TPP and its action uniformly among all possible choices; otherwise, the action that attains the highest output value of the network $Q(s, a; W)$ among all TPPs is selected. The probability ϵ decays with a decay rate of 0.99/episode along the training process, because we gain more and more confidence about the trained VTPN. Once an action is selected, the TPP is adjusted based on the action. After that, plan optimization is executed and DVHs of the plans before and after the TPP adjustment are used to compute the reward function r. The states s and s' prior to and after TPP adjustment, the chosen action a, and the reward r together form one training sample. By repeating this process, we are able to generate a large number of training samples. W is updated by the experience replay strategy to minimize the loss function in Eq. (5) via the stochastic gradient descent scheme each time with a batch of 16 randomly selected training samples. The purpose of the experience replay strategy is to overcome the strong correlation among sequentially generated training samples[21]. In this process, $\widehat{W}$ is updated by setting it to the latest network parameters W every 10 steps.

The VTPN framework is implemented using Python with TensorFlow[37] on a desktop workstation with eight Intel Xeon 3.5 GHz CPU processors, 32 GB memory and two Nvidia Quadro M4000 GPU cards.

2.5 Evaluations

We collect 74 patient cases with 10 patient cases for training and remaining 64 patient cases for testing. We evaluate the performance of the trained VTPN on both training and testing cases. For each case, we first set all TPPs to be unity. After that, the trained VTPN is used to operate the optimization engine in Eq. (1) in the workflow shown in Fig. 1. The iteration is terminated if VTPN determines to keep the TPPs unchanged at a step, or the number of adjustment steps reaches the maximum of 50. Among all the plans generated along the VTPN guided TPPs adjustment steps, we select the one with the highest plan score as the final plan. We compare the quality of the final plan with that of the initial plan generated using the initial TPPs. We evaluate plan quality using both the original and the modified ProKnow scores.

3. Results

3.1.1 Training results

The training process was successfully performed. After that, we applied the trained VTPN to training cases. Fig. 4 shows how the DVHs and dose distributions, TPP values, and plan scores evolve along the planning process in one training patient case. The final plan quality was substantially improved compared to the initial plan. The original and modified ProKnow scores showed increasing trends. These results illustrate that VTPN learnt the policy of how to operate the optimization engine to improve plan quality.

3.1.2 VTPN-based intelligent treatment planning process

We applied the trained VTPN to each of the 64 test patient cases that were not seen by VTPN in the training process. Fig. 5 shows the result of one representative case. The VTPN first decided to decrease the thresholding values of rectum in the first 6 steps to spare rectum dose. After that, it reduced $\tau_{bladder}$, which successfully decreased dose to bladder. Afterwards, VTPN determined to adjust weight factors of rectum and bladder, and finally focused on increasing λ to enhance the PTV homogeneity, until step 26 VTPN decided to keep all TPPs unchanged and the iteration could be terminated. In this process, the original ProKnow score increased from 5.34 to 8.70, which was close to the maximal score of 9. Meanwhile, the modified score increased from 6.19 to 11.71.

We report in Table 1 the performance of VTPN on all the training and testing patients. The VTPN can substantially improve the average score from 4.84 to 8.45 for the original ProKnow score and 6.07 to 10.93 for the modified score. On average, the planning process by VTPN takes ~3 mins, in which most of the time was spent to perform plan optimization after taking each TPPs adjustment.

In addition, we benchmark the performance of VTPN by comparing the planning result and efficiency of VTPN with a human physicist. With a good understanding of the optimization engine and extensive experience in treatment planning, the human physicist is able to achieve an average ProKnow score of ~8.5 on the testing cases, which is comparable to the performance of VTPN.

4. Discussion

The intelligent automatic treatment planning framework was developed under the motivation to model the intelligent behavior of human planners using VTPN. Different from most of the existing approaches adjusting TPPs by certain mathematical algorithms, VTPN was trained to spontaneously learn a policy that can intelligently tune the TPPs like a human planner. The training process via DRL allowed VTPN to explore different ways of TPP adjustments and memorize those good ones. This was similar to the human's learning process that generate intelligence in a trial-and-error way[21]. Although the number of training cases was very small, the actual data to train VTPN was the state-action pairs generated during the DRL process described in Sec. 2.4.3. The DRL process was able to generate a large amount of data, beneficial to train the large scale VTPN.

This study focused on a problem of prostate IMRT treatment planning for the consideration of the relatively simple problem with a small number of OARs. Yet, the proposed approach is generalizable to the treatment planning problem of other sites with more critical structures involved and other treatment modalities, such as volumetric modulated arc therapy (VMAT). The proposed framework has a rather generic structure that does not depend on the specific optimization problem of interest. Hence, it can be adopted to other treatment planning systems. For example, it may be coupled with the Eclipse treatment planning system through its Application Programming Interface (Varian Medical System, Palo Alto, CA), which is an ongoing project. Nevertheless, it is also admitted that generalization of the proposed method will encounter the difficulty of a larger number of TPPs to tune, which inevitably leads to a much larger VTPN and a higher computational burden.

The current study has the following limitations. First, the developed system is essentially a "black box" and it is difficult to interpret actions determined by VTPN. It is of central importance in deep learning to decipher the underlying intelligence[38–41]. It will be our future work to pursue this direction with human interpretable network techniques.

Second, the computational time needed for training VTPN is approximately a week on our desktop workstation under the current framework. The major computational efforts were spent to repetitively perform IMRT optimization along the DRL process, while the number of optimization needed is determined by the number of training cases, TPPs-tuning steps, and the training episodes. In this study, we restrict the number of training cases to be 10, while the maximal TPP-tuning steps and training episodes were set to be 25 and 100, respectively. Larger numbers should be applied to ensure the generalizability and robustness of VTPN. Furthermore, if we would like to move forward to handle more complicated treatment techniques, such as VMAT, the computational costs in plan optimization may grow significantly. In these regards, plan optimization efficiency needs to be significantly improved in future. One potential solution is to incorporate GPU-bases fast optimization algorithms in the training of VTPN.

In addition, the reward function based on the ProKnow score may not fully respect clinical criteria for plan quality evaluation. Refining and improving the existing plan quality evaluation criteria, e.g. adding DVHs of GTV and CTV and allowing hot spots in these structures to better fit clinical requirement is for sure one way to move forward. Moreover, an ideal reward function should also quantify physician's judgement. Recent advances in inverse deep reinforcement learning[23] allow learning the reward function based on human behaviors. In the treatment planning context, it may be possible to learn the physician's preference and incorporate it to train VTPN.

Moreover, the current VTPN judges plan quality only based on DVHs since DVH is a concise representation of optimized plan which human planner usually pays much attention to. As a preliminary study, we propose to focus on such simple representation to start with, and the results turned out to be encouraging. However, it is well known that DVH cannot capture position-specific information such as locations of hot/cold spots, which a physician often pays attention to. The complete information from a 3D dose image might be necessary

to more challenging planning tasks with more complicate TPS to operate. Further extending the system to include more clinically important features will be down the road.

5. Conclusions

In this paper, we proposed an intelligent automatic treatment planning framework empowered by a deep virtual treatment planner. Employing recent success in DRL, we demonstrate feasibility and potential of the proposed idea in the testbed of prostate cancer IMRT treatment planning. The developed VTPN was able to mimic the behaviors of a human planner to operate an optimization engine and generate satisfactory plans. To our knowledge, this was the first time that intelligent planning behaviors were autonomously generated through a training process. Our work showed the potential of using state-of-the-art deep learning techniques to create a virtual planner that can learn to master the challenging but clinically important task of radiotherapy treatment planning.

Acknowledgement

This work was supported by the National Institutes of Health grant number R01CA237269 and Cancer Prevention and Research Institute of Texas grant number RP160661.

References

1. I.M.R.T.C.W. Group, “Intensity-modulated radiotherapy: current status and issues of interest,” International Journal of Radiation Oncology* Biology* Physics 51, 880–914 (2001).
2. Oelfke U, Bortfeld T, “Inverse planning for photon and proton beams,” Medical dosimetry 26, 113–124 (2001). [PubMed: 11444513]
3. Webb S, “The physical basis of IMRT and inverse planning,” British Journal of Radiology 76, 678–689 (2003).
4. Xing L, Li JG, Donaldson S, Le QT, Boyer AL, “Optimization of importance factors in inverse planning,” Physics in Medicine and Biology 44, 2525 (1999). [PubMed: 10533926]
5. Lu R, Radke RJ, Happersett L, Yang J, Chui C-S, Yorke E, Jackson A, “Reduced-order parameter optimization for simplifying prostate IMRT planning,” Physics in Medicine & Biology 52, 849 (2007). [PubMed: 17228125]
6. Wu X, Zhu Y, “An optimization method for importance factors and beam weights based on genetic algorithms for radiotherapy treatment planning,” Physics in Medicine & Biology 46, 1085 (2001). [PubMed: 11324953]
7. Wang H, Dong P, Liu H, Xing L, “Development of an autonomous treatment planning strategy for radiation therapy with effective use of population-based prior data,” Medical Physics 44, 389–396 (2017). [PubMed: 28133746]
8. Yang Y, Xing L, “Inverse treatment planning with adaptively evolving voxel-dependent penalty scheme,” Medical Physics 31, 2839–2844 (2004). [PubMed: 15543792]
9. Wahl N, Bangert M, Kamerling CP, Ziegenhein P, Bol GH, Raaymakers BW, Oelfke U, “Physically constrained voxel-based penalty adaptation for ultra-fast IMRT planning,” Journal of Applied Clinical Medical Physics 17, 172–189 (2016). [PubMed: 27455484]
10. Yan H, Yin F-F, “Application of distance transformation on parameter optimization of inverse planning in intensity-modulated radiation therapy,” Journal of Applied Clinical Medical Physics 9, 30–45 (2008). [PubMed: 18714279]
11. Yan H, Yin F-F, Guan H.-q., Kim JH, “AI-guided parameter optimization in inverse treatment planning,” Physics in Medicine & Biology 48, 3565 (2003). [PubMed: 14653563]
12. Yan H, Yin F-F, Guan H, Kim JH, “Fuzzy logic guided inverse treatment planning,” Medical Physics 30, 2675–2685 (2003). [PubMed: 14596304]

13. Lee T, Hammad M, Chan TCY, Craig T, Sharpe MB, “Predicting objective function weights from patient anatomy in prostate IMRT treatment planning,” Medical Physics 40, 121706-n/a (2013). [PubMed: 24320492]
14. Boutilier JJ, Lee T, Craig T, Sharpe MB, Chan TCY, “Models for predicting objective function weights in prostate cancer IMRT,” Medical Physics 42, 1586–1595 (2015). [PubMed: 25832049]
15. Chan TCY, Craig T, Lee T, Sharpe MB, “Generalized Inverse Multiobjective Optimization with Application to Cancer Therapy,” Operations Research 62, 680–695 (2014).
16. LeCun Y, Bengio Y, Hinton G, “Deep learning,” Nature 521, 436–444 (2015). [PubMed: 26017442]
17. Chen L, Shen C, Li S, Maquilan G, Albuquerque K, Folkert MR, Wang J, presented at the Medical Imaging: Image Processing(2018).
18. Zhen X, Chen J, Zhong Z, Hrycushko B, Zhou L, Jiang S, Albuquerque K, Gu X, “Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study,” Physics in Medicine & Biology 62, 8246 (2017). [PubMed: 28914611]
19. Han YS, Yoo J, Ye JC, “Deep residual learning for compressed sensing CT reconstruction via persistent homology analysis,” arXiv preprint arXiv:1611.063912016).
20. Greenspan H, Van Ginneken B, Summers RM, “Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique,” IEEE Transactions on Medical Imaging 35, 1153–1159 (2016).
21. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, Petersen S, Beattie C, Sadik A, Antonoglou I, King H, Kumaran D, Wierstra D, Legg S, Hassabis D, “Human-level control through deep reinforcement learning,” Nature 518, 529 (2015). [PubMed: 25719670]
22. Shen C, Gonzalez Y, Chen L, Jiang S, Jia X, “Intelligent Parameter Tuning in Optimization-Based Iterative CT Reconstruction via Deep Reinforcement Learning,” IEEE transactions on medical imaging 37, 1430 (2018). [PubMed: 29870371]
23. Wulfmeier M, Ondruska P, Posner I, “Maximum entropy deep inverse reinforcement learning,” arXiv preprint arXiv:1507.048882015).
24. Iqbal Z, Nguyen D, Jiang S, “Super-Resolution 1H Magnetic Resonance Spectroscopic Imaging utilizing Deep Learning,” arXiv preprint arXiv:1802.079092018).
25. Ma G, Shen C, Jia X, “Low dose CT reconstruction assisted by an image manifold prior,” arXiv preprint arXiv:1810.122552018).
26. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D, “Mastering the game of Go with deep neural networks and tree search,” Nature 529, 484 (2016). [PubMed: 26819042]
27. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T, Hassabis D, “Mastering the game of Go without human knowledge,” Nature 550, 354 (2017). [PubMed: 29052630]
28. Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z, Jiang S, “Dose Prediction with U-net: A Feasibility Study for Predicting Dose Distributions from Contours using Deep Learning on Prostate IMRT Patients,” arXiv preprint arXiv:1709.092332017).
29. Nguyen D, Jia X, Sher D, Lin M-H, Iqbal Z, Liu H, Jiang S, “Three-Dimensional Radiotherapy Dose Prediction on Head and Neck Cancer Patients with a Hierarchically Densely Connected U-net Deep Learning Architecture,” arXiv preprint arXiv:1805.103972018).
30. Shen C, Gonzalez Y, Klages P, Qin N, Jung H, Chen L, Nguyen D, Jiang SB, Jia X, “Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer,” Physics in Medicine & Biology 64, 115013 (2019). [PubMed: 30978709]
31. Men C, Gu X, Choi D, Majumdar A, Zheng Z, Mueller K, Jiang S, “GPU-based ultra fast IMRT plan optimization,” Phys. Med. Biol 54, 6565–6573 (2009). [PubMed: 19826201]

32. Boyd S, Parikh N, Chu E, Peleato B, Eckstein J, "Distributed optimization and statistical learning via the alternating direction method of multipliers," Foundations and Trends® in Machine Learning 3, 1–122 (2011).
33. Glowinski R, Le Tallec P, Augmented Lagrangian and operator-splitting methods in nonlinear mechanics. (SIAM, 1989).
34. Watkins CJ, Dayan P, "Q-learning," Machine learning 8, 279–292 (1992).
35. Bellman R, Karush R, 1964.
36. LeCun Y, Bottou L, Bengio Y, Haffner P, "Gradient-based learning applied to document recognition," Proceedings of the IEEE 86, 2278–2324 (1998).
37. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, presented at the OSDI(2016).
38. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O, "Understanding deep learning requires rethinking generalization," arXiv preprint arXiv:1611.035302016).
39. Zhang Q, Wu YN, Zhu S-C, presented at the The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)(2018).
40. Che Z, Purushotham S, Khemani R, Liu Y, presented at the AMIA Annual Symposium Proceedings(2016).
41. Sturm I, Lapuschkin S, Samek W, Müller K-R, "Interpretable deep neural networks for single-trial EEG classification," Journal of neuroscience methods 274, 141–145 (2016). [PubMed: 27746229]

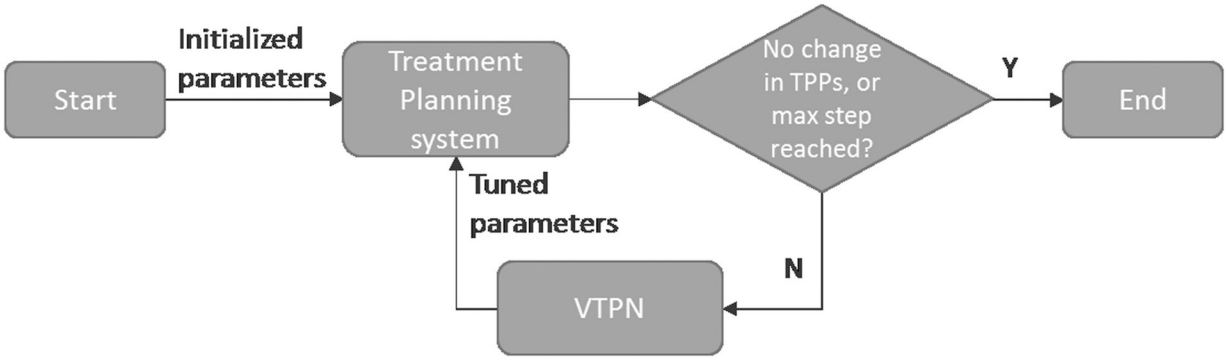

Figure 1.
Overall workflow of the proposed intelligent automatic treatment planning process.

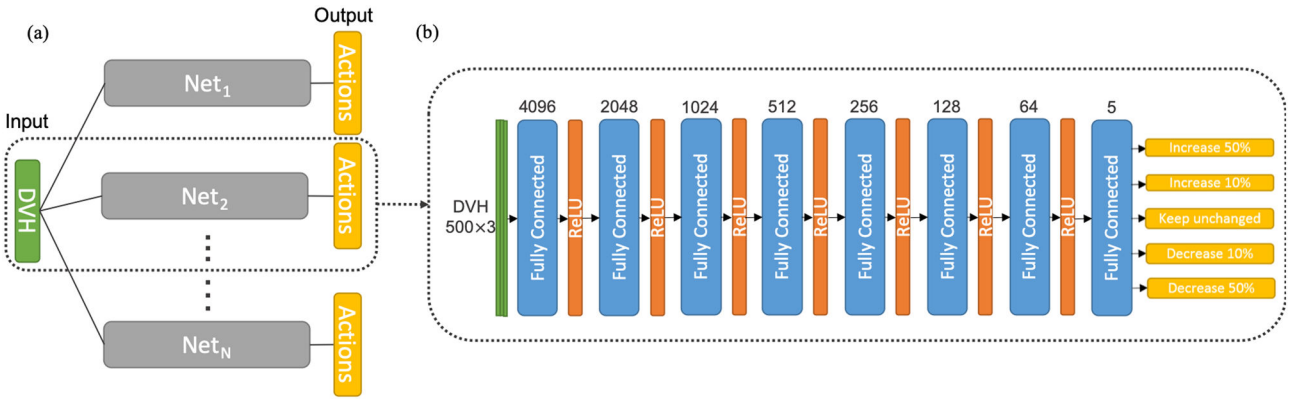

Figure 2.

Network structure of the VTPN. (a) The overall structure of VTPN. (b) Detailed structure of a subnetwork. Size after each fully connected layer is specified at the top. The output layer of VTPN gives the Q-value associated with each TPPs adjustment action. Each action corresponds to an output of the final fully-connected layer.

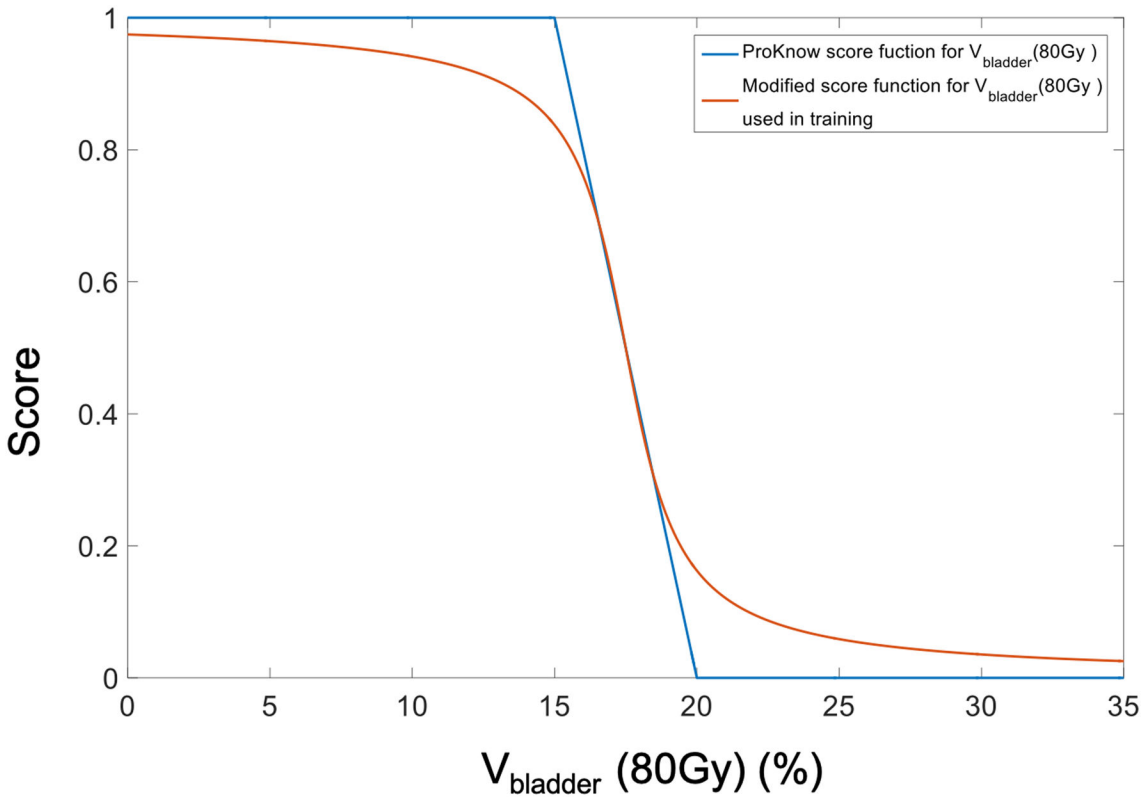

Figure 3.
ProKnow score function (blue curve) for $V_{bladder}$ (80Gy) and modified score function used in training (red curve).

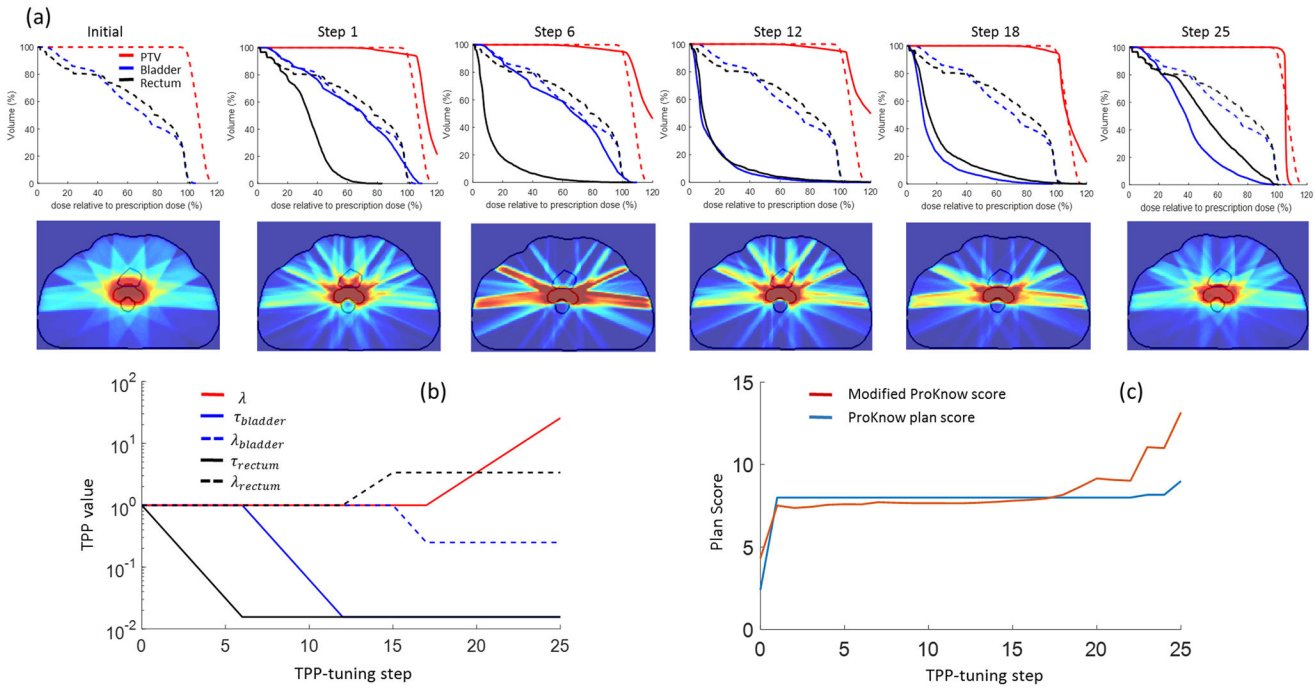

Figure 4.

Evolution of DVHs and dose distributions (a), TPP values (b) and original and modified ProKnow plan scores (c) in the planning process of a training patient case.

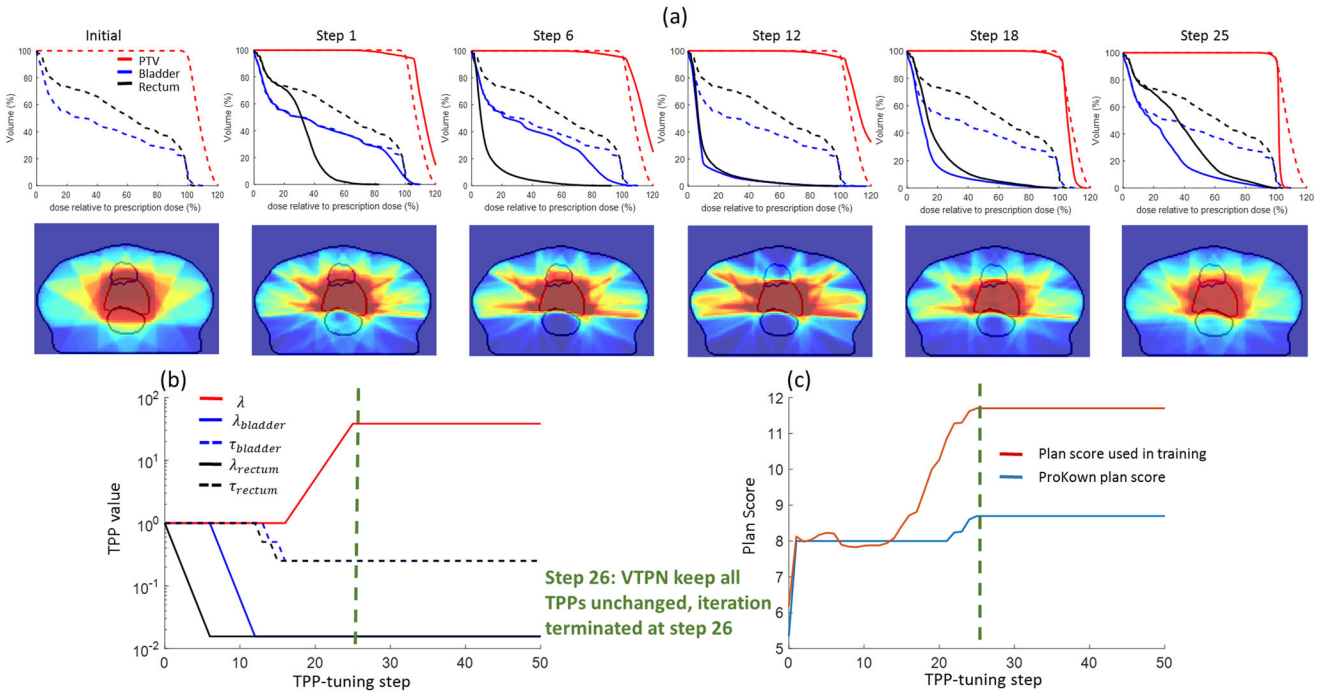

Figure 5.

Evolution of DVHs and dose distributions (a), TPP values (b) and original and modified ProKnow scores (c) in the planning process of a testing patient case.

Table 1.

Average scores (±standard deviation) of plans generated by VTPN compared to the initial plan with all TPPs set to unity.

	ProKnow plan score (Initial)	ProKnow plan score (VTPN-guided)	Modified score (Initial)	Modified score (VTPN-guided)
Training	4.04 (±2.36)	8.46 (±0.50)	5.33 (±2.12)	10.85 (±2.22)
Testing	4.97 (±2.02)	8.45 (±0.48)	6.19 (±1.94)	10.94 (±2.08)
Overall	4.84 (±2.07)	8.45 (±0.48)	6.07 (±1.97)	10.93 (±2.09)