IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 14, 2019, accepted October 17, 2019, date of publication October 28, 2019, date of current version November 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2949871

Aperture Shape Generation Based on Gradient Descent With Momentum

LIYUAN ZHANG[1], PENGCHENG ZHANG[1], JIE YANG[2], JIE LI[3], AND ZHIGUO GUI[1]

[1]Shanxi Provincial Key Laboratory for Biomedical Imaging and Big Data, North University of China, Taiyuan 030051, China
[2]School of Medicine Management, Shanxi University of Chinese Medicine, Taiyuan 030619, China
[3]Department of Radiation Oncology, Shanxi Provincial Cancer Hospital, Taiyuan 030013, China

Corresponding author: Zhiguo Gui (gzgtg@163.com)

This work was supported by in part by the National Natural Science Foundation of China under Grant 11605160, in part by the Research Project supported by the Shanxi Scholarship Council of China under Grant 2016-089, in part by the National Key Scientific Instrument and Equipment Development Project of China under Grant 2014YQ24044508, in part by the National Natural Science Foundation of China under Grant 61671413, and in part by the National Key Research and Development Program of China under Grant 2016YFC0101605.

ABSTRACT Direct aperture optimization (DAO) is an effective method to generate high-quality intensity-modulated radiation therapy treatment plans. In generic DAO, the direction of negative gradient descent is generally used to determine the aperture shape. However, this strategy can reduce the convergence rate, especially near the optimal value. We propose aperture shape generation based on the direction of gradient descent with momentum, where column generation is implemented as carrier. During aperture shape generation of column generation, the current aperture gradient map is first calculated. Then, the gradient with momentum is calculated based on the existing gradient information. Finally, the direction of gradient descent with momentum is constructed for obtaining the deliverable aperture shape by solving the pricing problem. To verify the effectiveness of the proposed method, we conducted comparative experiments on two head and neck and two prostate tumor cases. Compared with generic column generation, the proposed method can effectively protect the organs at risk while ensuring the required dose distribution to the target. Using the proposed method, the number of apertures and optimization time can be reduced by up to 30.95 and 32.96%, respectively, compared to the conventional approach. The experimental results suggest that the proposed method can accelerate the search speed and improve the quality of treatment plans.

INDEX TERMS Aperture shape, column generation, direct aperture optimization, gradient descent direction, gradient descent with momentum.

I. INTRODUCTION

Direct aperture optimization (DAO) [1]–[4] usually considers three approaches: stochastic search [5]–[8], local gradient-based method [9], and column generation [10]–[13]. Although the optimization strategies of these three approaches differ, aperture shape generation consists of selecting an appropriate gradient descent direction for the improvement of the objective function. Many studies are available on aperture shape generation. The genetic algorithm to optimize aperture shape has been proposed in [7], [8]. In addition, simulated annealing based on the gradient has been employed for optimization [14], and subsequently, aperture shape optimizations based on fuzzy enhancement [15] and region growth [16] have been proposed. To generate the aperture shape, these methods directly use the aperture gradient map, that is, they determine the aperture shape by constructing the negative gradient descent direction. The negative gradient descent method enables optimization with an inexpensive computation and can converge to a local minimum from any initial solution. However, when approaching the minimum, the method is prone to the sawtooth phenomenon, that is, both the step length per iteration and the convergence speed reduce. To overcome slow convergence, we propose an improved aperture shape generation method based on the idea of gradient descent with momentum [17], [18] to accelerate optimization and improve the quality of the treatment plan. During aperture shape generation, as the gradient vectors across iterations do not always point in the same direction, the aperture gradients can be regarded as noisy. Consequently, search may fall into a local optimum. The momentum method in deep learning resembles momentum properties in physics for accelerating

The associate editor coordinating the review of this manuscript and approving it for publication was Jiafeng Xie.

convergence speed and leaving local optima [19]. Gradient descent with momentum accelerates the search progress by accumulating the existing gradient information and continuing to move along the leading direction.

DAO based on stochastic search can avoid local minima with a certain probability. However, as the search is based on local gradient information, it cannot guarantee that the optimization result is the global optimum. In addition, the local gradient-based method only optimizes the aperture shape according to this local information and depends on a good initial solution, thus not guaranteeing global optimality. Unlike these methods, column generation starts with an empty aperture set, and there is no initial solution. Then, it constructs a network flow with global gradient information to obtain the deliverable aperture shape by solving the pricing problem. Hence, column generation provides a strict theoretical derivation and adopts a complete search strategy to ensure global optimality, being the most accurate method compared to stochastic search and local gradient-based methods.

We propose aperture shape generation based on the direction of gradient descent with momentum, where column generation is used as carrier for implementation. During column generation, the proposed method first calculates the aperture gradient map. Then, according to the momentum and current gradient, the direction of gradient descent with momentum is determined. Finally, the pricing problem uses the direction of gradient descent with momentum to obtain the deliverable aperture shape. We evaluate the proposed method in two head and neck and two prostate tumor cases, and compare the results with those obtained from generic column generation to verify the effectiveness and performance of the proposed method. The rest of this study is organized as follows. Section II introduces the proposed method. The experimental settings, evaluation criteria, objective function and results are detailed in section III. Section IV discusses the results. Finally, we draw conclusions in section V.

II. METHODS

In this section, we introduce the proposed aperture shape generation. Section II-A details the calculation of the dose received from the aperture. Section II-B summarizes generic column generation, and section II-C details the gradient descent with momentum. Finally, section II-D discusses the proposed aperture shape generation based on gradient descent with momentum.

A. DOSE CALCULATION

Each beam applied to the patient can be decomposed into a set of beamlets labeled as B in a rectangular grid of m rows and n columns, where the size of a beamlet is 1×1 cm^2. The set of all deliverable apertures is denoted as K, where the aperture weight is y_k ($k \in K$). A_k represents the set of beamlets exposed by multi-leaf collimator (MLC) in aperture k, S represents the total number of organs and tissues of the current patient, and v_s represents the number of voxels in structure s ($s \subset S$). The dose received by voxel j ($j = 1, \cdots, v_s$) in structure s at unit intensity from beamlet i of aperture A_k is the corresponding element, W_{ijs}, in the deposition matrix (i.e., deposition coefficient). Thus, dose D_{js} received by voxel j in structure s ($s \subset S$) can be expressed as

$$D_{js} = \sum_{k \in K} \left(\sum_{i \in A_k} W_{ijs} \right) y_k, \quad j = 1, \cdots, v_s, \ s = 1, \cdots, S. \tag{1}$$

B. COLUMN GENERATION

During plan optimization for step-and-shoot intensity-modulated radiation therapy (IMRT), the optimization problem is expressed as

$$minmize\ F(D) = minmize \sum_{s=1}^{S} \sum_{\xi=1}^{N_s} F_{\xi s}(D_s), \tag{2}$$

subject to

$$\sum_{k \in K} \left(\sum_{i \in A_k} W_{ijs} \right) y_k = D_{js}, \quad j = 1, \cdots, v_s, \ s = 1, \cdots, S, \tag{3}$$

$$y_k > 0, \quad k \in K, \tag{4}$$

where, in objective function $F(D)$, $F_{\xi s}(D_s)$ is the ξ-th subobjective function of structure s, and D_s is the dose distribution of that structure. N_s subobjective functions are used to control the dose distribution received by structure s. In each iteration of column generation to generate the treatment plan, the pricing problem is firstly solved to obtain the deliverable aperture. Then, in the master problem, a gradient-based method optimizes the weights of all the obtained apertures. The iterative process continues until the optimization result meets the clinical requirements or the number of iterations reaches its limit, and the treatment plan is retrieved.

Column generation can generate aperture by solving the pricing problem, which is derived from the optimization problem of IMRT plan optimization. Under inequality constraints, the Lagrange function is first constructed for optimization, and then the KKT (Karush–Kuhn–Tucker) conditions are obtained to finally determine the optimal solution. The Lagrange function considering (2)–(4) is given by

$$\begin{aligned} L(D_{js}, y_k, \pi_{js}, \rho_k) = & \sum_{s=1}^{S} \sum_{j=1}^{v_s} F_{js}(D_{js}) \\ & + \sum_{k=1}^{K} \rho_k(-y_k) + \sum_{s=1}^{S} \sum_{j=1}^{v_s} \pi_{js} \\ & \times \left(\sum_{k \in K} \left(\sum_{i \in A_k} W_{ijs} \right) y_k - D_{js} \right), \end{aligned} \tag{5}$$

where, ρ_k and π_{js} are Lagrange multipliers. When the optimal solution is obtained, the KKT conditions to be satisfied are

$$\nabla_{D_{js}, y_k} L\left(D_{js}, y_k, \pi_{js}, \rho_k\right) = 0, \tag{6}$$

$$-y_k \leq 0, \quad k \in K, \tag{7}$$

$$\sum_{k\in K}\left(\sum_{i\in A_k} W_{ijs}\right) y_k - D_{js} = 0, \quad j = 1, \cdots, v_s, \ s = 1, \cdots, S, \tag{8}$$

$$\rho_k \geq 0, \quad k \in K, \tag{9}$$

$$-\rho_k y_k = 0, \quad k \in K. \tag{10}$$

The Lagrange multipliers can be obtained from the above KKT conditions as

$$\pi_{js} = \frac{\partial F_{js}(D_{js})}{\partial D_{js}}, \tag{11}$$

$$\rho_k = \sum_{s=1}^{S}\sum_{j=1}^{v_s}\left(\sum_{i\in A_k} W_{ijs}\right)\pi_{js}. \tag{12}$$

From (9) and (12), we obtain

$$\sum_{s=1}^{S}\sum_{j=1}^{v_s}\left(\sum_{i\in A_k} W_{ijs}\right)\pi_{js} > 0, \tag{13}$$

that is, the optimal solution should satisfy

$$\min_{k\in K}\sum_{i\in A_k}\left(\sum_{s=1}^{S}\sum_{j=1}^{v_s} W_{ijs}\pi_{js}\right) > 0. \tag{14}$$

If the current solution satisfies (14), it is the optimal one for the corresponding plan. The current generated aperture is not added to the plan. Therefore, column generation to obtain the treatment plan should determine the pricing of each deliverable aperture:

$$\min_{k\in K}\sum_{i\in A_k}\left(\sum_{s=1}^{S}\sum_{j=1}^{v_s} W_{ijs}\pi_{js}\right). \tag{15}$$

Then, the pricing problem can be converted into a shortest path problem by using the network flow constructed according to the aperture gradient map [20].

C. GRADIENT DESCENT WITH MOMENTUM

In this study, the mechanical constraints of the MLC system for step-and-shoot IMRT involved do not allow interdigitation. To determine the aperture shape during column generation with that MLC mechanical constraint, we use the network flow to solve the pricing problem [20]. In generic column generation, the gradients corresponding to the beamlets is directly used to determine the aperture shape and select the optimal descent direction for the objective function. This descent direction is similar to that selected by the steepest descent method. In the steepest descent method, the update step of the method is $x_{k+1} = x_k - d_k$, where d_k is the update from independent variable x_k to x_{k+1} (i.e., $d_k = \alpha_k * \nabla f_k$), α_k is the step length, and ∇f_k is the first derivative of objective function $f(x)$ at x_k. From any initial solution, the steepest descent method can find the local minimum of the objective function. However, when the vertical variation is faster than the horizontal one in the contour of the objective function, that is, the vertical gradient of two consecutive instants is reversed, the steepest descent method changes the search direction frequently and can produce oscillation while approaching the local minimum. In deep learning [21], the method of gradient descent with momentum considers the previous descent directions and accumulates the velocity of the gradient motion. When the gradient direction is consistent in an update, it moves faster in this direction, reducing oscillations during search.

In gradient descent with momentum, the update step of the method is also $x_{k+1} = x_k - d_k^M$, but update d_k^M of independent variable x_k to x_{k+1} is the weighted vector sum of the current gradient descent and the last update, that is, the sum of $d_k = \alpha_k * \nabla f_k$, and the update d_{k-1}^M from x_{k-1} to x_k is multiplied by coefficient β_k in [0, 1]:

$$d_k^M = d_k + d_{k-1}^M * \beta_k = \alpha_k * \nabla f_k + d_{k-1}^M * \beta_k, \tag{16}$$

where current gradient descent direction d_k is the basic search direction, and last update d_{k-1}^M is the auxiliary search direction.

If the vector angle is $\theta_k \in [0, \pi/2)$ between gradient descent direction d_k and last update d_{k-1}^M (Fig. 1(a)), d_{k-1}^M contributes a positive acceleration to search direction d_k. In contrast (Fig.1(b)), if the vector angle is $\theta_k \in (\pi/2, \pi]$, d_{k-1}^M decelerates d_k. When d_{k-1}^M is orthogonal to d_k, we consider d_{k-1}^M to have no effect on d_k. The influence of d_{k-1}^M on d_k can be adjusted by parameter β_k. Compared with the steepest descent method, the method of gradient descent with momentum not only reduces oscillations, but also preserves the general search direction aiming to ensure efficiency and a correct convergence to the optimal solution.

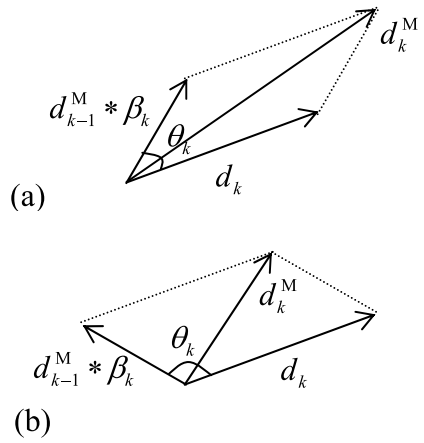


FIGURE 1. The effect of momentum gradient descent direction to original gradient descent direction (a) acceleration effect; (b) deceleration effect.

D. PROPOSED APERTURE SHAPE GENERATION

Gradient descent with momentum is specially suited for noisy gradients. To solve the pricing problem, the gradient in the aperture gradient map, which is used to construct the network flow and generate the aperture, can be regarded as noisy.

Therefore, we adopt gradient descent with momentum to modulate the aperture gradient map and efficiently generate the deliverable treatment plan. In an iteration of column generation to obtain the treatment plan, the aperture shape generation updates gradient sequence g_k to construct gradient map $G_k(m, n)$ according to the current information, and then it constructs the network flow to solve the pricing problem and generate the deliverable aperture shape [16], [20]. If gradient sequence g_k is directly used to solve the pricing problem, the aperture shape generation uses the negative gradient descent direction to determine the shape. If the pricing problem is solved by using the gradient sequence considering gradient descent with momentum, the aperture shape is determined correspondingly. Inspired by the gradient descent with momentum, when calculating the gradient sequence according to the available information, the accumulated gradient information, g_{k-1}^M, from the previous iterations is introduced into the calculation of current gradient information g_k with a certain weight, and the coefficients are normalized to obtain gradient sequence g_k^M containing the accumulated gradient information g_{k-1}^M and current gradient information g_k, obtaining (17), as shown at the bottom of this page. Variable $g_k(i)$ is the i-th gradient element of the gradient sequence obtained from the information of the objective function and generated aperture during the k-th iteration. Gradient element $g_k^M(i)$ corresponding to $g_k(i)$ is modulated considering the momentum, and $g_{k-1}^M(i)$ is the modulated gradient element during the $(k-1)$-th iteration. If angle $\theta_k \in \left[0, \pi/2\right)$ exists between $g_k(i)$ and $g_{k-1}^M(i)$ (i.e., $g_k(i) \cdot g_{k-1}^M(i) > 0$), $g_{k-1}^M(i)$ is introduced into $g_k^M(i)$ with a larger weight α for $g_{k-1}^M(i)$ to accelerate $g_k(i)$ in the positive direction. Otherwise (i.e., $g_k(i) \cdot g_{k-1}^M(i) < 0$), $g_{k-1}^M(i)$ is introduced into $g_k^M(i)$ with a small weight $(1-\alpha)$ to exert a deceleration on $g_k(i)$. If $g_k(i)$ is orthogonal to $g_{k-1}^M(i)$ (i.e., $g_k(i) \cdot g_{k-1}^M(i) = 0$), $g_k(i)$ remains unchanged. In particular, at the optimal value, $g_k(i) = 0$ for any i, and (6) holds. Therefore, at the optimal value obtained by the proposed method, according to (17), there is $g_k^M(i) = g_k(i) = 0$ for any i, and (6) still holds. Hence, the proposed method does not change the optimal solution. The above strategy is adopted to modulate the gradient sequence, and then the network flow is constructed to solve the pricing problem [20] while speeding up the search and improving the plan quality.

III. EXPERIMENTS AND RESULTS

We evaluated two cases of head and neck tumors and two cases of prostate tumors to experimentally verify the proposed aperture shape generation based on gradient descent with momentum in comparison to generic column generation. The dose deposition matrix, W, to obtain the dose was obtained by the classical pencil beam dose calculation method [22] in the Computational Environment for Radiological Research open-source software [23]. Different types of subobjective functions were used to form the total objective function, including minimum dose, maximum dose, mean dose, and dose-volume (DV) criterion subobjective functions [24], to control the dose coverage on each structure. During master problem solving of the proposed and comparison methods, the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm for bound constrained optimization [25]–[27] was used to optimize the weight of each aperture. We denote the proposed aperture shape optimization based on gradient descent with momentum as M, and generic column generation as CG.

A. EXPERIMENTAL SETTINGS

As mentioned above, two cases of head and neck tumors and two cases of prostate tumors were used to verify the effectiveness of the proposed method compared with generic column generation.

For the head and neck tumor cases shown in Fig. 2(a), nine 6 MeV co-irradiated photon fields at intervals of 40° were used to irradiate the target. Three targets were considered, labeled as planning target volume (PTV), PTV 70 Gy, PTV 63 Gy, and PTV 56 Gy. In addition, two parotid glands (ipsilateral parotid gland—IL-PG and contralateral parotid gland—CL-PG), brain stem, and spinal cord were selected as the organs at risk (OARs) [28]. The remaining tissue is denoted as Tissue. As shown in Fig. 2(b), for the prostate tumor cases, five 6 MeV co-irradiated photon fields were used to irradiate the target at frame angles of 36, 100, 180, 260, and 324°. For these cases, only one target was labeled as PTV. The bladder and rectum were selected as the OARs, and the rest of tissues were denoted as Tissue. Among the four tumor cases, the two head and neck tumor cases were labeled as H1 and H2, and the two prostate tumor cases were labeled as P1 and P2.

B. EVALUATION CRITERIA

To evaluate the quality of the optimized plans obtained by the two evaluated methods, the dose-volume histogram (DVH) of the OARs were first evaluated according to the clinical guidelines by Marks *et al.* [29] on the premise of ensuring dose distribution to the target (see Table 1). Then, the conformity

$$g_k^M(i) = \begin{cases} \dfrac{\left((2-\alpha)\, g_k(i) + \alpha g_{k-1}^M(i)\right)}{2} & \text{if } \left(g_k(i) \cdot g_{k-1}^M(i) > 0\right) \\ g_k(i) & \text{if } \left(g_k(i) \cdot g_{k-1}^M(i) = 0\right) \\ \dfrac{\left((1+\alpha)\, g_k(i) + (1-\alpha)\, g_{k-1}^M(i)\right)}{2} & \text{if } \left(g_k(i) \cdot g_{k-1}^M(i) < 0\right) \end{cases}, \quad \alpha \in (0.5, 1). \tag{17}$$

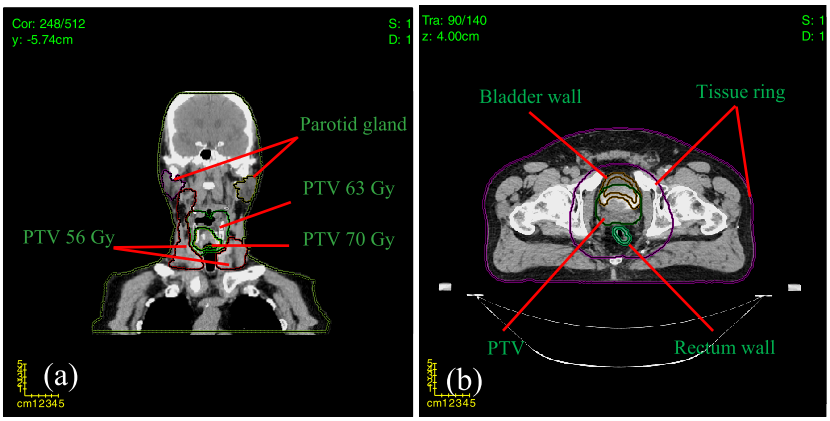


FIGURE 2. Structural distribution of various organs (a) the head and neck tumor case; (b) the prostate tumor case.

TABLE 1. DV constraint conditions of OARs.

Head and neck Cases		Prostate Cases		
Structure	DV parameter	Structure	Bladder	Rectum
Parotid gland	$D_{mean} \leq 25$ Gy	DV parameter	$V_{65\,Gy} < 50\%$ $V_{70\,Gy} < 35\%$ $V_{75\,Gy} < 25\%$ $V_{80\,Gy} < 15\%$	$V_{50\,Gy} < 50\%$ $V_{60\,Gy} < 35\%$ $V_{65\,Gy} < 25\%$ $V_{70\,Gy} < 20\%$ $V_{75\,Gy} < 15\%$
Spinal cord	$D_{max} \leq 50$ Gy			
Brain stem	$D_{max} \leq 54$ Gy			

number (CN) [30] and homogeneity index (HI) [31] of the target were respectively calculated as

$$CN = \frac{TV_{ri}}{TV} \times \frac{TV_{ri}}{V_{ri}}, \tag{18}$$

$$HI = \frac{D_{5\%}}{D_{95\%}}, \tag{19}$$

where TV is the total volume of the target, TV_{ri} is the volume of the target within the 95% isodose line, V_{ri} is the volume of all tissues within the 95% isodose line, and $D_{5\%}$ and $D_{95\%}$ are the radiation doses received by 5% and 95% of the target volume, respectively. Values of CN and HI close to 1 indicate a more suitable dose distribution to the target.

The generalized equivalent uniform dose (gEUD) and normal tissue complication probability (NTCP) were calculated to evaluate the protective effect of the optimization method to the OARs. In the head and neck tumor cases, the radiobiological parameters of the NTCP model for the parotid gland were retrieved from [32], and those for the spinal cord and brain stem from [33]. In the prostate tumor cases, the radiobiological parameters of the NTCP model for the bladder wall were retrieved from [34], and those for the rectum wall from [35]. In addition, as there is no uniform international standard for biological evaluation criteria such as NTCP and gEUD, we considered that lower NTCP and gEUD values indicate higher protection to the OARs. Finally, the running time and number of apertures of the evaluated methods were determined.

C. OBJECTIVE FUNCTION

The objective function was a linear combination of weighted subobjective functions [36] and expressed as

$$f\left(D\left(x\right)\right) = \sum_{i=1}^{I} \xi_i f_{Type}\left(D^{Structure}\left(x\right)\right), \tag{20}$$

where the dose distribution $D\left(x\right)$ is a linear function of W and the fluence matrix x; that is, $D\left(x\right) = Wx$. I is the number of subobjective functions, ξ_i is the weight factor corresponding to the i-th subobjective function, $D^{Structure}\left(x\right)$ is the dose distribution of each structure, and $f_{Type}\left(D^{Structure}\left(x\right)\right)$ denotes the different types subobjective functions. For the head and neck tumor cases, the DV criterion subobjective function [24] was used to constrain dose distribution of the parotid gland, the maximum dose subobjective function was used to punish the dose in the spinal cord and brain stem that exceeds the maximum dose of the clinical guideline [29], the mean dose and minimum dose subobjective functions were used to constrain the dose distribution of the three targets, and the maximum dose subobjective function to constraint the dose distribution of Tissue. For the prostate tumor cases, the DV criterion subobjective function proposed by Wu and Mohan [24] was used to constrain the dose distribution of the

TABLE 2. The optimization information of head and neck tumor cases.

		H1		H2	
		CG	M	CG	M
PTV 70 Gy	$V_{70\,Gy}$ (%)	97.7157	99.2786	99.3109	98.8584
	$V_{77\,Gy}$ (%)	0	0	0	0
	HI	1.0486	1.0397	1.0426	1.0471
	CN	0.8885	0.09571	0.8609	0.7345
PTV 63 Gy	$V_{63\,Gy}$ (%)	97.77	98.5064	96.2442	95.0236
	$V_{70\,Gy}$ (%)	0.6317	3.0363	2.5653	4.3116
	HI	1.0836	1.0906	1.0926	1.1087
	CN	0.2149	0.1293	0.0159	0.0133
PTV 56 Gy	$V_{56\,Gy}$ (%)	97.6525	97.9484	96.9753	96.8676
	$V_{62\,Gy}$ (%)	0.47693	1.0634	4.0094	3.5226
	HI	1.0666	1.0704	1.0888	1.0882
	CN	0.0023	0.0036	4.9858E-05	3.2194E-05
IL–PG	Mean dose (Gy)	22.5747	21.0398	24.6433	24.8871
	gEUD (Gy)	22.5745	21.0398	24.6436	24.8873
	NTCP (%)	12.72	7.50	23.12	24.60
CL–PG	Mean dose (Gy)	16.5337	11.2133	21.9502	20.8968
	gEUD (Gy)	16.5338	11.2132	21.9502	20.8968
	NTCP (%)	1.01	3.87E-02	10.35	7.11
Spinal cord	Max dose (Gy)	49.475	45.625	49.225	49.925
	gEUD (Gy)	41.5341	37.9123	41..2295	40.8559
	NTCP (%)	1.60	0.70	1.49	1.38
Brain stem	Max dose (Gy)	29.575	29.325	49.775	48.525
	gEUD (Gy)	13.5102	14.7405	34.2437	31.8914
	NTCP (%)	7.65E-07	1.67E-06	3.63E-02	1.37E-02
Aperture number		90	69	84	58
Time (s)		927.998	622.146	1026.417	698.105

bladder and rectum, the mean dose and minimum dose subobjective functions were used to constrain the dose distribution of the target, and the DV criterion subobjective function to constrain the dose distribution of Tissue.

D. RESULTS

We first verified the proposed method for cases H1 and H2. Generic column generation for comparison considered a maximum number of apertures of 100. The proposed method was varied to obtain similar optimization results as the comparison method, and we registered the number of required apertures. The corresponding results are shown in Fig. 3 and Table 2. Compared to generic column generation, the proposed method can reduce the dose on the OARs, as shown by the reduced NTCP, and ensure the dose distribution to multiple targets. In addition, the proposed method can generate similar quality plans with fewer apertures. Moreover, the proposed method can accelerate the DAO convergence by using gradient descent with momentum to determine the aperture shape.

When using generic column generation to optimize the plans for cases P1 and P2, the maximum number of apertures was set to 60. Like for cases H1 and H2, we aimed to obtain similar results for both methods and registered the number of required apertures from the proposed method. The corresponding results are shown in Fig. 4 and Table 3. Compared to generic column generation, the dose distribution to that target from the proposed method is guaranteed, and both the NTCP and gEUD of the OARs decrease. Therefore, the optimization results from the two prostate tumor cases further verify that the proposed method can retrieve similar or better treatment plans with fewer apertures compared to generic column generation.

IV. DISCUSSION

To evaluate the experimental results, the curves of the target in the DVH should be similar for the compared methods, even the target DVH obtained by the proposed method outperforms that obtained by generic column generation. On this basis, the curves of the OARs in the DVH obtained from the compared methods were further evaluated. The DVHs of case H1 are shown in Figs. 3(a) and (b). The DVH of multiple targets in Fig. 3(b) and the dose distribution in Table 2 show that compared with generic column generation, the proposed method ensures a consistent and even improved dose distribution to the targets. These results show that the curves of the OARs in the DVH obtained by the proposed method in Fig. 3(a) are substantially lower than those obtained by generic column generation. According to clinical guidelines [29], the mean dose in the parotid gland should be below 25 Gy, and the dose of the spinal cord and brain stem should not exceed 50 and 54 Gy, respectively. From the optimization results for H1 in Table 2, we can conclude that, compared to generic column generation, the mean doses on the two parotid glands obtained by the proposed method notably reduce, as can be confirmed from the dose distribution in Figs. 3(e) and (f). Furthermore, the maximum doses at the spinal cord and brain stem obtained by the proposed

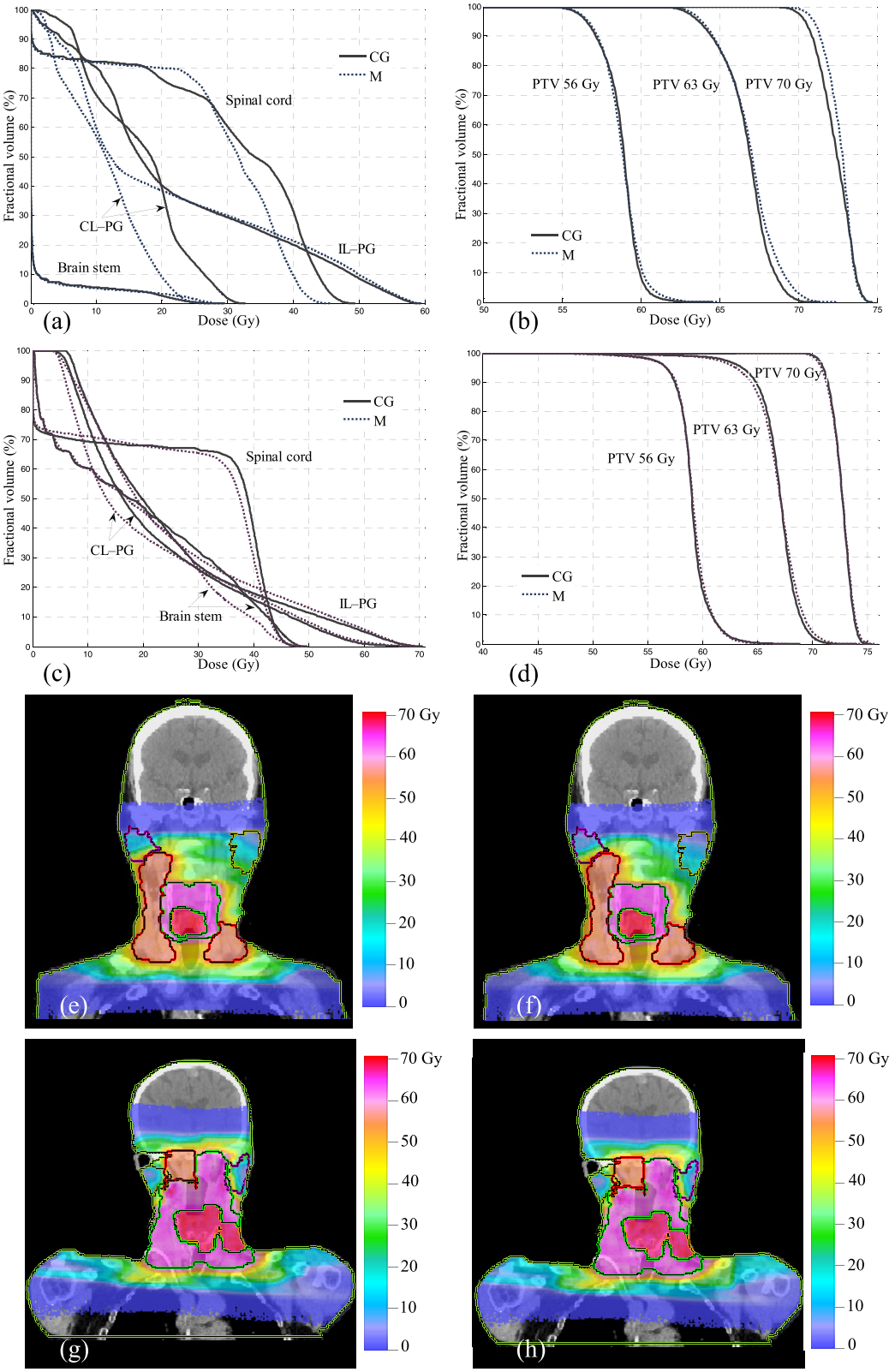


FIGURE 3. Comparison of optimization results of head and neck tumor cases (a) DVH of OARs in H1; (b) DVH of targets in H1; (c) DVH of OARs in H2; (d) DVH of targets in H2; (e) the dose distribution of H1 optimized by generic column generation; (f) the dose distribution of H1 optimized by the proposed method; (g) the dose distribution of H2 optimized by generic column generation; (h) the dose distribution of H2 optimized by the proposed method.

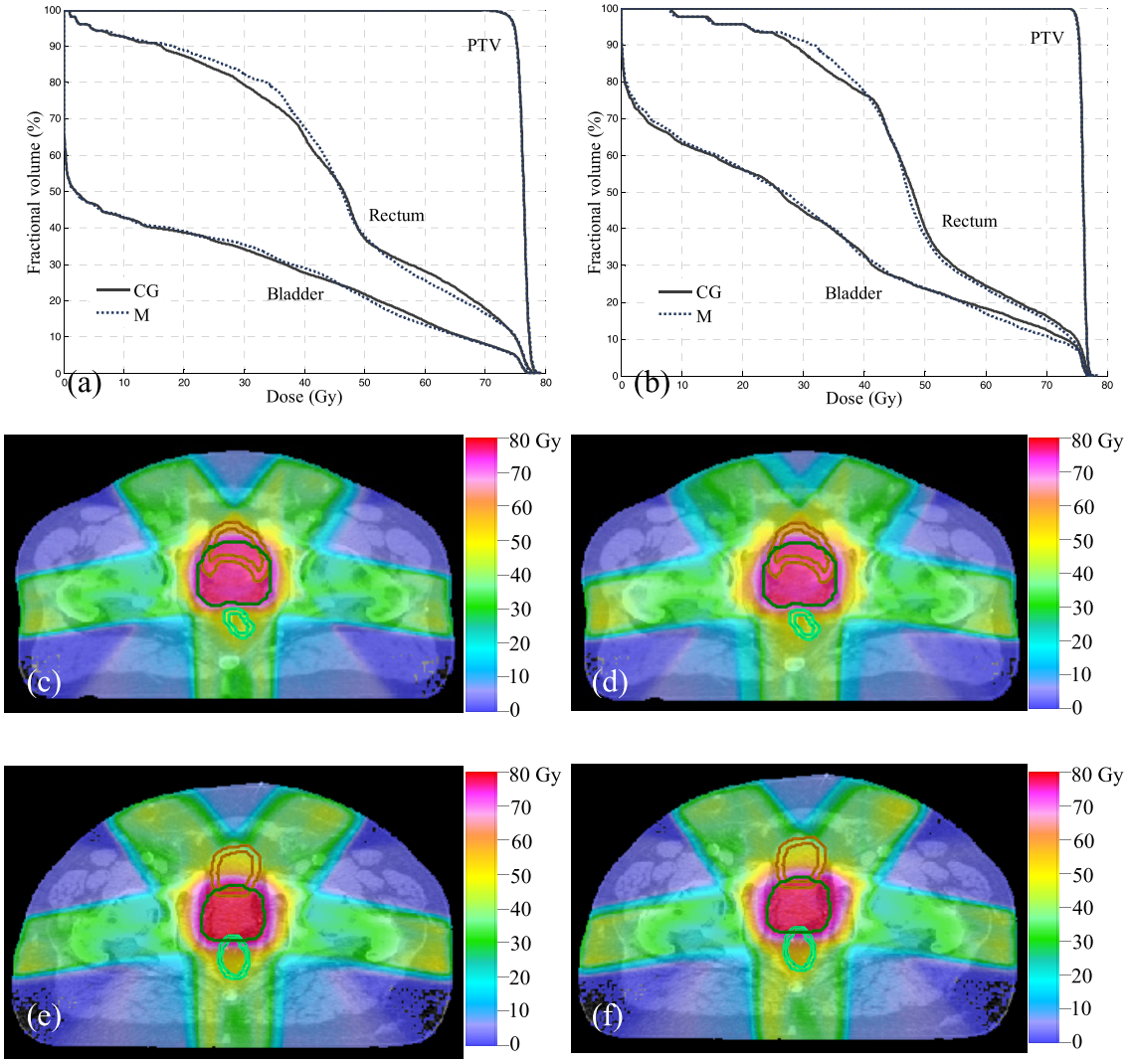


FIGURE 4. Comparison of optimization results of prostate tumor cases (a) DVH of P1; (b) DVH of P2; (c) the dose distribution of P1 optimized by generic column generation; (d) the dose distribution of P1 optimized by the proposed method; (e) the dose distribution of P2 optimized by generic column generation; (f) the dose distribution of P2 optimized by the proposed method.

method comply with clinical guidelines [29], being slightly lower than those obtained from generic column generation. In addition, the number of apertures of the proposed method decreases by 23.33%, and the optimization time decreases by 32.96% with respect to generic column generation, showing an improved optimization performance from the proposed method.

The DVHs of case H2 are shown in Figs. 3(c) and (d). Along with the information in Table 2, the curve of PTV 56 Gy in the DVH obtained from the proposed method is slightly better than that obtained from generic column generation, whereas the curves of PTV 63 and 70 Gy in the DVH meet the clinical requirements and are close to the results obtained from generic column generation. On this basis, the curve of the CL-PG in the DVH obtained from the proposed method in Fig. 3(c) decreases compared with that obtained from generic column generation, as seen in the dose distributions of Figs. 3(g) and (h). The optimization results for case H2 in Table 2 show that, compared with the results obtained from generic column generation, the mean dose on the CL-PG reduces in the proposed method. In addition, the mean dose of IL-PG obtained by the two methods comply with clinical guidelines [29]. The maximum doses of the spinal cord and brain stem comply with the dose constraint, and the gEUD and NTCP values are lower in the proposed method than in generic column generation. Moreover, the number of apertures from the proposed method to generate similar optimization results compared to generic column generation reduces by 30.95%, and the optimization time reduces by 31.99%. Note that the locations and morphology of the tumors substantially vary in the two cases of head and neck tumors, as shown in Figs. 3(e) and (g). Although the improved performance of the proposed method compared to generic column generation is not substantial for

TABLE 3. The optimization information of prostate tumor cases.

		P1		P2	
		CG	M	CG	M
PTV	$V_{67.27\,Gy}$ (%)	99.9779	99.9985	100	100
	$V_{74\,Gy}$ (%)	98.1124	98.1249	99.8292	99.853
	$V_{81.4\,Gy}$ (%)	0	0	0	0
	HI	1.0359	1.0379	1.0212	1.0223
	CN	0.8576	0.8295	0.8976	0.9042
Bladder	gEUD (Gy)	56.8275	56.5628	59.2063	58.5605
	NTCP (%)	4.94	4.74	7.03	6.40
Rectum	gEUD (Gy)	62.3378	61.8932	61.768	61.2272
	NTCP (%)	7.57	7.01	6.86	6.23
Aperture number		55	47	57	47
Time (s)		669.120	520.873	775.751	545.160

case H2, the results verify its suitability to achieve similar or better optimization results with fewer apertures in a shorter optimization time than the comparison method.

Figs. 4(c) and (e) show the locations and morphology of organs for prostate tumor cases P1 and P2. In Figs. 4(a) and (b), the curves of the target in the DVH in these cases mostly coincide among the two evaluated methods. The clinical guidelines in [29] specify that the DV constraints of the OARs (i.e., bladder and rectum) in prostate tumor cases are concentrated on the part of high dose. On this basis, in Figs. 4(a) and (b), compared with generic column generation, the high-dose part of DVH in the bladder and rectum for cases P1 and P2 obtained from the proposed method decreases. Considering the information in Table 3, both the NTCP and gEUD of the bladder and rectum from the proposed method are lower than those from generic column generation. The number of apertures decreases by 14.55% (17.54%), and the optimization time decreases by 22.16% (29.72%) for case P1 (P2) when optimized by the proposed method with respect to generic column generation. Hence, the results for the prostate tumor cases verify that the proposed method can achieve similar or better optimization results with less apertures and in shorter time compared to generic column generation. Overall, experimental results show that the proposed method can generate treatment plans meeting clinical guidelines [29].

During aperture modulation by gradient descent with momentum, α can be adjusted to change the effect of the cumulative momentum on the current gradient calculation. Generally, the value of α in (17) is heuristically determined. In this study, through several experiments, we determined that when $\alpha = 0.99$, the experimental effect is more notable.

When multiple targets are optimized, the proposed method may not notably improve the HI of targets and cannot guarantee improvement in their CN. Hence, we will aim to ensure such improvements in future work. In addition, as the proposed method accelerates generic column generation via software, we will aim to introduce the concept of gradient descent with momentum into other DAO algorithms and combine the proposed method with hardware acceleration for faster generation of treatment plans complying clinical requirements.

V. CONCLUSION

Generic aperture shape generation considers the negative gradient descent direction to determine the aperture shape, but convergence to the solution may be slow. To improve the convergence speed, we generate the aperture shape based on gradient descent with momentum, where column generation is used as carrier. The proposed method leverages the concept of gradient descent with momentum to modulate the gradient information and weighs the previous gradient information into the current gradient calculation to speed up search and improve the plan quality. Experimental results show that the proposed method can ensure the dose distribution to the target while protecting the OARs, accelerating the optimization process, and shortening the optimization time. The proposed method was suitably applied to different tumor cases, suggesting its feasibility for clinical application.

REFERENCES

[1] K. Otto, "Volumetric modulated arc therapy: IMRT in a single gantry arc," *Med. Phys.*, vol. 35, no. 1, pp. 310–317, Jan. 2008, doi: 10.1118/1.2818738.

[2] J. Unkelbach, T. Bortfeld, D. Craft, M. Alber, M. Bangert, R. Bokrantz, D. Chen, R. J. Li, L. Xing, C. H. Men, S. Nill, D. Papp, E. Romeijn, and E. Salari, "Optimization approaches to volumetric modulated arc therapy planning," *Med. Phys.*, vol. 42, no. 3, pp. 1367–1377, Mar. 2015, doi: 10.1118/1.4908224.

[3] M. A. Earl, D. M. Shepard, S. Naqvi, X. A. Li, and C. X. Yu, "Inverse planning for intensity-modulated arc therapy using direct aperture optimization," *Phys. Med. Biol.*, vol. 48, no. 8, pp. 1075–1089, Apr. 2003, doi: 10.1088/0031-9155/48/8/309.

[4] K. Bzdusek, H. Friberger, K. Eriksson, B. Hårdemark, D. Robinson, and M. Kaus, "Development and evaluation of an efficient approach to volumetric arc therapy planning," *Med. Phys.*, vol. 36, no. 6, pp. 2328–2339, Jun. 2009, doi: 10.1118/1.3132234.

[5] D. M. Shepard, M. A. Earl, X. A. Li, S. Naqvi, and C. Yu, "Direct aperture optimization: A turnkey solution for step-and-shoot IMRT," *Med. Phys.*, vol. 29, no. 6, pp. 1007–1018, Jun. 2002, doi: 10.1118/1.1477415.

[6] M. A. Earl, M. K. N. Afghan, C. X. Yu, Z. Jiang, and D. M. Shepard, "Jaws-only IMRT using direct aperture optimization," *Med. Phys.*, vol. 34, no. 1, pp. 307–314, Jan. 2007, doi: 10.1118/1.2403966.

[7] Y. Li, J. Yao, and D. Yao, "Genetic algorithm based deliverable segments optimization for static intensity-modulated radiotherapy," *Phys. Med. Biol.*, vol. 48, no. 20, pp. 3353–3374, Oct. 2003, doi: 10.1088/0031-9155/48/20/007.

[8] C. Cotrutz and L. Xing, "Segment-based dose optimization using a genetic algorithm," *Phys. Med. Biol.*, vol. 48, no. 18, pp. 2987–2998, Sep. 2003, doi: 10.1088/0031-9155/48/18/303.

[9] B. Hardemark, A. Liander, H. Rehbinder, and J. Lof, "Direct machine parameter optimization with raymachine in pinnacle," RaySearch Lab., Stockholm, Sweden, White Paper WP-DMPO rev. 1, 0310, 2003.

[10] F. Carlsson, "Combining segment generation with direct step-and-shoot optimization in intensity-modulated radiation therapy," *Med. Phys.*, vol. 35, no. 9, pp. 3828–3838, Sep. 2008, doi: 10.1118/1.2964096.

[11] F. Preciado-Walters, M. P. Langer, R. L. Rardin, and V. Thai, "Column generation for IMRT cancer therapy optimization with implementable segments," *Ann. Oper. Res.*, vol. 148, no. 1, pp. 65–79, Nov. 2006, doi: 10.1007/s10479-006-0080-1.

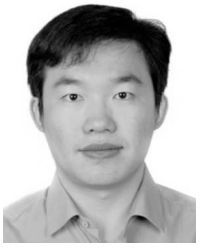
[12] C. Men, H. E. Romeijn, Z. C. Taşkin, and J. F. Dempsey, "An exact approach to direct aperture optimization in IMRT treatment planning," *Phys. Med. Biol.*, vol. 52, no. 24, pp. 7333–7352, Dec. 2007, doi: 10.1088/0031-9155/52/24/009.

[13] C. Men, X. Jia, and S. B. Jiang, "GPU-based ultra-fast direct aperture optimization for online adaptive radiation therapy," *Phys. Med. Biol.*, vol. 55, no. 15, pp. 4309–4319, Aug. 2010, doi: 10.1088/0031-9155/55/15/008.

[14] J. Yang, P. Zhang, L. Zhang, and Z. Gui, "A gradient-based direct aperture optimization," (in Chinese), *J. Biomed. Eng.*, vol. 35, no. 3, pp. 358–367, 2018.

[15] P. Zhang, L. Zhang, J. Yang, and Z. Gui, "The aperture shape optimization based on fuzzy enhancement," *IEEE Access*, vol. 6, pp. 35979–35987, 2018, doi: 10.1109/ACCESS.2018.2849208.
[16] L. Zhang, Z. Gui, J. Yang, and P. Zhang, "A column generation approach based on region growth," *IEEE Access*, vol. 7, pp. 31123–31139, 2019, doi: 10.1109/ACCESS.2019.2896175.
[17] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1–17, Dec. 1964, doi: 10.1016/0041-5553(64)90137-5.
[18] N. Qian, "On the momentum term in gradient descent learning algorithms," *Neural Netw.*, vol. 12, no. 1, pp. 145–151, 1999, doi: 10.1016/S0893-6080(98)00116-6.
[19] G. Ma, "Research on the key image processing problems of X-ray detection for ultra large-scale integrated circuit packaging," Ph.D. dissertation, South China Univ. Technol. North Campus, Guangzhou, China, 2016, pp. 27–28.
[20] H. E. Romeijn, R. K. Ahuja, J. F. Dempsey, and A. Kumar, "A column generation approach to radiation therapy treatment planning using aperture modulation," *SIAM J Optim.*, vol. 15, no. 3, pp. 838–862, 2005, doi: 10.1137/040606612.
[21] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006, doi: 10.1162/neco.2006.18.7.1527.
[22] A. Ahnesjö, M. Saxner, and A. A. Trepp, "A pencil beam model for photon dose calculation.," *Med. Phys.*, vol. 19, no. 2, pp. 263–273, Mar./Apr. 1992, doi: 10.1118/1.596856.
[23] J. O. Deasy, A. I. Blanco, and V. H. Clark, "CERR: A computational environment for radiotherapy research," *Med. Phys.*, vol. 30, no. 5, pp. 979–985, May 2003, doi: 10.1118/1.1568978.
[24] Q. Wu and R. Mohan, "Algorithms and functionality of an intensity modulated radiotherapy optimization system," *Med. Phys.*, vol. 27, no. 4, pp. 701–711, Apr. 2000, doi: 10.1118/1.598932.
[25] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM J. Sci. Comput.*, vol. 16, no. 5, pp. 1190–1208, 1995, doi: 10.1137/0916069.
[26] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal, "Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization," *ACM Trans. Math. Softw.*, vol. 23, no. 4, pp. 550–560, 1997, doi: 10.1145/279232.279236.
[27] J. L. Morales and J. Nocedal, "Remark on 'Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization,'" *ACM Trans. Math. Softw.*, vol. 38, no. 1, Nov. 2011, Art. no. 7, doi: 10.1145/2049662.2049669.
[28] G. Mu, E. Ludlum, and P. Xia, "Impact of MLC leaf position errors on simple and complex IMRT plans for head and neck cancer," *Phys. Med. Biol.*, vol. 53, no. 1, pp. 77–88, Jan. 2008, doi: 10.1088/0031-9155/53/1/005.
[29] L. B. Marks, E. D. Yorke, A. Jackson, R. K. Ten Haken, L. S. Constine, A. Eisbruch, S. M. Bentzen, J. Nam, and J. O. Deasy, "Use of normal tissue complication probability models in the clinic," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 76, no. 3, pp. S10–S19, Mar. 2010, doi: 10.1016/j.ijrobp.2009.07.1754.
[30] A. van't Riet, A. C. Mak, M. A. Moerland, L. H. Elders, and W. van der Zee, "A conformation number to quantify the degree of conformality in brachytherapy and external beam irradiation: Application to the prostate," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 37, no. 3, pp. 731–736, Feb. 1997, doi: 10.1016/S0360-3016(96)00601-3.
[31] N. Hodapp, "The ICRU report 83: Prescribing, recording and reporting photon-beam intensity-modulated radiation therapy (IMRT)," *Strahlentherapie Onkologia*, vol. 188, no. 1, pp. 97–99, Jan. 2012, doi: 10.1007/s00066-011-0015-x.
[32] A. Eisbruch, R. K. Ten Haken, H. M. Kim, L. H. Marsh, and J. A. Ship, "Dose, volume, and function relationships in parotid salivary glands following conformal and intensity-modulated irradiation of head and neck cancer," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 45, no. 3, pp. 577–587, Oct. 1999, doi: 10.1016/S0360-3016(99)00247-3.
[33] C. Burman, G. J. Kutcher, B. Emami, and M. Goitein, "Fitting of normal tissue tolerance data to an analytic function," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 21, no. 1, pp. 123–135, May 1991, doi: 10.1016/0360-3016(91)90172-Z.
[34] E. Dale, T. P. Hellebust, A. Skjønsberg, T. Hogberg, and D. R. Olsen, "Modeling normal tissue complication probability from repetitive computed tomography scans during fractionated high-dose-rate brachytherapy and external beam radiotherapy of the uterine cervix," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 47, no. 4, pp. 963–971, Jul. 2000, doi: 10.1016/S0360-3016(00)00510-1.
[35] S. T. Peeters, M. S. Hoogeman, W. D. Heemsbergen, A. A. Hart, P. C. Koper, and J. V. Lebesque, "Rectal bleeding, fecal incontinence, and high stool frequency after conformal radiotherapy for prostate cancer: Normal tissue complication probability modeling," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 66, no. 1, pp. 11–19, Sep. 2006, doi: 10.1016/j.ijrobp.2006.03.034.
[36] M. L. Kessler, D. L. Mcshan, M. A. Epelman, K. A. Vineberg, A. Eisbruch, T. S. Lawrence, and B. A. A. Fraass "Costlets: A generalized approach to cost functions for automated optimization of IMRT treatment plans," *Optim. Eng.*, vol. 6, no. 4, pp. 421–448, Dec. 2005, doi: 10.1007/s11081-005-2066-2.

LIYUAN ZHANG is currently pursuing the Ph.D. degree with the North University of China, with a focus on the research direction is the optimization of the plan of precision radiotherapy.



PENGCHENG ZHANG received the Ph.D. degree in computer science and technology from Southeast University, Nanjing, China, and the Ph.D. degree in traitement du signal et tél écommunications from the Université de Rennes 1, Rennes, France, in 2014. He currently engages in teaching and research with the North University of China. His research interests include medical image reconstruction, medical image analysis, dose calculation, and planning optimization.

JIE YANG received the B.Sc. and M.Sc. degrees in computer science and technology, in 2004 and 2009, respectively, and the Ph.D. degree in information and communication engineering from the North University of China, in 2018. Her research interests include programming, dose calculation, and planning optimization.

JIE LI received the master's degree from Shanxi Medical University, in 2009. He is currently the Chief Physician with the Department of Radiation Oncology, Shanxi Provincial Cancer Hospital, engaged in cancer radiotherapy work.

ZHIGUO GUI received the Ph.D. degree in signal and information processing from the North University of China, in 2004. He is currently a Professor with the North University of China. His research interests include image processing and image reconstruction.