Leveraging Visual Prompt with Diffusion Adversarial Network for Radiotherapy Dose Prediction

Zhenghao Feng[1], Lu Wen[1], Jiaqi Cui[1], Xi Wu[2], Jianghong Xiao[3], Xingchen Peng[4], Dinggang Shen[5,6,7] and Yan Wang[1(✉)]

[1] School of Computer Science, Sichuan University, China
wangyanscu@hotmail.com
[2] School of Computer Science, Chengdu University of Information Technology, China
[3] Department of Radiation Oncology, Cancer Center, West China Hospital, Sichuan University, China
[4] Department of Biotherapy, Cancer Center, West China Hospital, Sichuan University, China
[5] School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, Shanghai Tech University, China
[6] Shanghai United Imaging Intelligence Co., Ltd., China
[7] Shanghai Clinical Research and Trial Center, China

Abstract. Automatic prediction of dose distribution maps wields considerable influence in clinical radiotherapy treatment. Recently, deep learning-based approaches have been explored to automatically predict the dose map from structure images and obtain promising results. However, these methods mainly focus on extracting anatomical features from CT and organ masks, ignoring abundant visual knowledge inherent in the domain of dose map. To address this limitation, we innovatively propose a visual prompt-guided dose prediction model, named ViPDose, to effectively predict radiotherapy dose distribution for cancer patients. Specifically, our ViPDose is structured with two key stages: 1) a prompt pre-training stage and 2) a prompt generation stage. In the pre-training stage, we train a prompt encoder to encode dose maps alongside structure im-ages into compact prompt vectors. Then, in the prompt generation stage, we design a fast prompt generator fulfilled with a diffusion adversarial network (DAN) to efficiently produce the prompt vectors that closely approximate those generated by the prompt encoder, thus enriching the model with abundant visual prompt information. By adopting DAN in such highly compressed latent space, our method can guarantee high-quality predictions with relatively low computation costs. Comprehensive experiments on a clinical rectal cancer dataset with 130 cases have verified the superior performance of our method over other state-of-the-art methods.

Keywords: Radiotherapy, Dose Prediction, Diffusion Adversarial Network.

Z. Feng and L. Wen—The authors contribute equally to this work.

1 Introduction

Radiotherapy serves as a crucial non-surgical treatment for fighting cancers [1]. To ensure curative potency, a clinically acceptable radiotherapy plan needs to exert a high-energy prescription dose on planning target volume (PTV) while sparing surrounding organs at risk (OARs) from radiation-induced harm. To reach this, the dosimetrists have to engage in iterative clinical trials and manual parameter optimization through a cumbersome trail-and-error process, unavoidably causing treatment delays [2]. Besides, potential differences in expertise among dosimetrists may result in undesirable curative inconsistency. Consequently, automatic dose prediction techniques have garnered significant attention for expediting and standardizing the plan-making procedures [3-4].

Nowadays, with the spectacular developments of deep learning (DL) in various medical fields [5-8], DL-based dose prediction methods have been introduced to automatically predict the dose distribution maps from several structure images, i.e., Computed Tomography (CT) scans and segmentation masks of PTV and OARs. The predicted dose distribution can help the physicians to derive objectives parameters with less optimization iterations in treatment planning system (TPS), finally gaining executable configurations. These methods are generally divided into two categories: regression-based methods [9-15] and generative-based methods [16-18]. For the regression-based ones, UNet [19] has been widely employed with various innovative modules for achieving dose prediction tasks [10-13]. For example, Tan et al. [10] introduced two related tasks, i.e., isodose line prediction and gradient map prediction, to assist the main dose prediction task to gain higher dose accuracy. Wang et al. [12] proposed the PRUNET to progressively refine the dose prediction of prostate cancer. Jiao et al. [9] introduced the GCN into dose prediction task. Despite their promising accuracy, these regression-based methods confront unavoidable blurry predictions with insufficient high-frequency information which may reveal the ray directions and dose attenuation [18].

To maintain the high-frequency details, generative-based methods have emerged as effective solutions. Based on generative adversarial networks (GANs) [21], complex image statistics can be implicitly learned, thus gaining perceptually realistic predictions with more sharp patterns [16-17]. More recently, the diffusion model (DM) [22] stands out as another powerful generative model that does not rely on any additional assumptions about target data distribution [23-25]. In this vein, Feng et al. [18] presented a diffusion-based dose prediction method (DiffDP) and used an iterative denoising process to produce high-quality dose maps. Nevertheless, due to the significant difference between input images and the output dose maps, existing DL-based methods [9-19] mainly focus on extracting anatomical features only from the input CT image and organ masks while overlooking the important knowledge in dose map. Such knowledge from dose map is of great potential to boost the prediction network learning the nonlinear relationship between the structure images and final dose distribution. Besides, the enhancement of current DM-based methods (e.g., DiffDP) come at the cost of heavy computational burden due to the multiple denoising iterations and high input resolution [26] which seriously restricts the practical applications of DMs in dose prediction tasks.

In this paper, we innovatively propose **Vi**sual **P**rompt-guided **Dose** prediction model with a diffusion adversarial network (DAN), named **ViPDose**, to achieve accurate dose

distribution for radiotherapy. Unlike most previous methods that rely solely on patients' structure images and attempt to learn the direct mapping from structure images to dose maps without any prior of their underlying data distribution, our method incorporates knowledge in dose map as visual prompt to facilitate more reliable prediction. Specifically, the whole framework contains a prompt pretraining stage and a prompt generation stage. Different from merely using the real dose map to constrain the predicted one in the output level, we design the prompt pretraining stage and train a prompt encoder to directly compress dose map (i.e., ground truth) along with the structure images into compact prompt vectors. Such prompt vectors can guide the main prediction network to reach a more accurate dose prediction. However, such essential prompt information extracted from the real dose map is inaccessible in the inference, so we design the prompt generation stage to train a prompt generator to produce such prompts. Concretely, we fix the parameters of the prompt encoder and the prediction network, and then train a DAN, which is modified from DDGAN [27], to efficiently produce the prompt vector as approximated to that from prompt encoder as possible, thus explicitly learning the data distribution of the dose maps. Finally, in the inference, input with the structure image, the DAN generates the corresponding prompt vector through a reverse process and guides the prediction network to predict the dose maps for cancer patients. Different from traditional DMs, ViPDose has two advantages: (i) ViPDose uses a modified DDGAN [27] as prompt generator where adversarial training is used to enlarge step size and enhance both the training and inference process; (ii) ViPDose is devised to the highly compact latent space, notably reducing the computational burdens with smaller input resolution.

In summary, the main contributions of this work can be concluded into three-fold as follows: (1) We present a novel visual prompt-guided dose prediction model to fully explore the prompt knowledge about dose distribution to reach effective dose prediction of rectal cancer. (2) We innovatively design a fast prompt generator with adversarial training to model the data distribution of the prompt vector in the latent space, ensuring the prediction quality with a much lower computational cost. (3) Extensive experiments on a clinical dataset with 130 rectal cancer patients have confirmed the superior performance of our method compared to other state-of-the-art (SOTA) approaches.

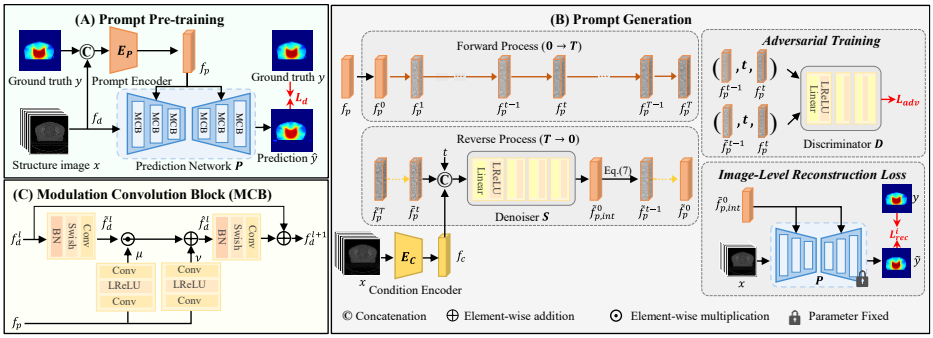

Fig. 1. Illustration of the proposed ViPDose framework.

2 Methodology

The overview of ViPDose is illustrated in Fig. 1 which contains (A) a prompt pretraining stage and (B) a prompt generation stage. Stage (A) trains a prompt encoder E_p to effectively encode the structure images and dose map (i.e., ground truth) into compressed prompt vector which can guide the prediction network P to complete dose map generation. Then, Stage (B) trains a DDGAN only conditioned on structure images to generate prompt vector in the latent space and guide the dose map reconstruction. For the effectiveness of prompt generation, the feature-level and image-level reconstruction constraints are designed to associate with the adversarial loss for network optimization.

2.1 Prompt Pre-training

Prompt Encoder. To extract the beneficial prompt knowledge of the dose map, we design a prompt pre-training stage, as depicted in Fig. 1(A), to train a prompt encoder E_p for compressing the dose map (i.e., ground truth) into compact feature vectors. We denote an image set of a patient as $\{x, y\}$ where structure images $x = (x_{CT}, x_{SM})$ contains CT $x_{CT} \in \mathbb{R}^{1\times H\times W}$ (H for height and W for width) and binary masks of PTV and OARs $x_{SM} \in \mathbb{R}^{(O+1)\times H\times W}$ (O is the number of OARs), and $y \in \mathbb{R}^{1\times H\times W}$ is the ground truth. Later, we concatenate y with structure images x along the channel dimension which is then fed into E_p to form a prompt vector $f_p \in \mathbb{R}^{C'}$ (C' for channel numbers).

Prediction Network. We design a five-level U-Net-like prediction network P for predicting the dose map $\tilde{y}.$, where the conventional convolution blocks are replaced by modulation convolution blocks (MCBs). Specifically, as displayed in Fig. 1(C), MCB fuses f_p with the dose feature f_d (extracted by the prediction network from the structure image x) in multiple levels. For the l-th encoding or decoding level, MCB uses two Conv-LeakyReLU-Conv groups to map f_p into two modulation parameters, i.e., μ and ν, which share the same channel numbers with dose feature f_d^l. Subsequently, μ and ν are integrated with the dose feature f_d^l for gaining the fused feature $\hat{f}_d^l$ with scaling and shifting transform. This process can be formally described as below:

$$\hat{f}_d^l = \tilde{f}_d^l \odot \mu + \nu, \tag{1}$$

$$\tilde{f}_d^l = Conv(Swish(BN(f_d^l))), \tag{2}$$

where $Conv(\cdot)$, $Swish(\cdot)$, and $BN(\cdot)$ are the convolution layer, swish activation function, and batch normalization layer, respectively. With spatial dimensions maintained, MCB can obtain both spatial-wise transformation and feature-wise manipulation. Then, we feed $\hat{f}_d^l$ into another Conv-Swish-BN group and add it to the original f_d^l with residual connection, gaining the dose feature f_d^{l+1} for the next encoding or decoding level:

$$f_d^{l+1} = Conv\left(Swish\left(BN\left(\hat{f}_d^l\right)\right)\right) + f_d^l, \tag{3}$$

To guarantee the effectiveness of the prompt encoder E_p in extracting prompt vectors, we optimize E_p and the prediction network P with the dose loss L_d which evaluates the difference between the prediction $\tilde{y}$ and its ground truth y as follows:

$$L_d = \|y - \tilde{y}\|_1, \tag{4}$$

2.2 Prompt Generation

After the prompt pre-training stage, the well-trained prompt encoder E_P and prediction network P are obtained. Since the prompt vector is extracted from the real dose map which is inaccessible in the inference, we design a fast prompt generator to synthesize the prompt merely under the condition of structure images. Inspired by DDGAN [27], the prompt generator, i.e., DAN, incorporates DM and adversarial training to enlarge the step size of DM. To force the prompt generator to synthesize more reliable prompts, dual-level reconstruction constraints are devised to better optimize the whole network.

Diffusion Framework. Given the structure images x and dose map y, the pre-trained and parameter-fixed E_p converts them into the prompt vector $f_p \in \mathbb{R}^{C'}$. Then, we utilize a DAN to generate the prompt vector to be approximated to f_p. Consistent with DM, the diffusion framework of DAN involves a forward process and a reverse process. The *forward process* produces a series of noisy prompt vectors $\{f_p^0, f_p^1, \dots, f_p^T\}$, f_p^0=f_p by progressively exerting a small amount of noise to f_p in T steps with the noise increased at each step until it becomes a standard Gaussian noise at step T [22]. Given $f_p^0 \sim q(f_p^0)$, the forward process is formulated as:

$$q\left(f_p^t \mid f_p^{t-1}\right) = \mathcal{N}\left(f_p^t; \sqrt{\alpha_t} f_p^{t-1}, (1-\alpha_t)I\right), t = 1, 2, \dots, T, \tag{5}$$

where $\alpha_{1:T}$ is the constant variance schedule of the noise added to f_p^{t-1}. Denoting $\gamma_t = \prod_{i=1}^{t} \alpha_i$, f_p^t at any time step t can be sampled by:

$$f_p^t = \sqrt{\gamma_t} f_p^0 + \sqrt{1-\gamma_t}\varepsilon_t, \varepsilon_t \sim \mathcal{N}(0, I) \tag{6}$$

where $\mathcal{N}(0, I)$ represents the standard Gaussian distribution. In this procedure, the generator G, i.e., the composite function of conditional encoder E_c and denoiser S, is trained to predict an intermediate prompt vector $\tilde{f}_{p,int}^0 = G(f_p^t, t, x)$ for the corresponding step t, under the guidance of structure images x.

The *reverse process* gradually converts the latent distribution $p(f_p^T)$ into $p(f_p^0)$ with the well-trained G, under the guidance of x. Beginning with a Gaussian distribution $f_p^T \sim \mathcal{N}(f_p^T|0, I)$, the reverse inference between two adjacent steps is expressed as:

$$\tilde{f}_p^{t-1} = \sqrt{\gamma_{t-1}} \tilde{f}_{p,int}^0 + \sqrt{1-\gamma_{t-1}} z_{t-1}, z_{t-1} \sim \mathcal{N}(0, I) \tag{7}$$

where $\tilde{f}_{p,int}^0 = G(f_p^t, t, x)$. Our parameterization of directly predicting $\tilde{f}_{p,int}^0$ is not only equivalent to predicting ε_t in vanilla DMs [18], but also simplifies the following adversarial training. Besides, our framework is applied to the highly compact feature level where its dimension is much smaller than conventional image-level DMs [18], thus effectively reducing the computational cost and enhancing the inference speed.

Adversarial Training. Inspired by DDGAN [27], we modified the original diffusion framework into an adversarial version. The forward process in Eq. (5) requires a relatively large T and small step size to satisfy the Gaussian assumption on the denoising distribution, which results in severely low inference efficiency. When timesteps T

decreases, the real denoising distribution $q\left(f_p^{t-1} \mid f_p^t\right)$ turns to a more complex non-Gaussian distribution. Therefore, we introduce adversarial training to model the complex data distribution between f_p^{t-1} and f_p^t given a large step size t.

In the adversarial training, the generator G generates the predicted intermediate prompt vector $\tilde{f}_{p,int}^0$ which is transformed into $\tilde{f}_p^{t-1}$ through Eq. (7), while the discriminator D tries to discriminate whether the samples come from $q\left(f_p^{t-1} \mid f_p^t\right)$ or $q\left(\tilde{f}_p^{t-1} \mid f_p^t\right)$. The adversarial loss L_{adv} can be formulated as below:

$$L_{adv} = L_{adv}(G) + L_{adv}(D), \tag{8}$$

$$L_{adv}(G) = \mathbb{E}_{q\left(\tilde{f}_p^{t-1} \mid f_p^t\right)}\left[-log\left(D\left(\tilde{f}_p^{t-1}, t, f_p^t\right)\right)\right], \tag{9}$$

$$\begin{aligned} L_{adv}(D) = \mathbb{E}_{q\left(f_p^{t-1} \mid f_p^t\right)}\left[-log\left(D\left(f_p^{t-1}, t, f_p^t\right)\right)\right] \\ + \mathbb{E}_{q\left(\tilde{f}_p^{t-1} \mid f_p^t\right)}\left[-log\left(1 - D\left(\tilde{f}_p^{t-1}, t, f_p^t\right)\right)\right], \end{aligned} \tag{10}$$

Conditional Encoder and Denoiser. To guide the denoising procedure with anatomical knowledge, a conditional encoder E_c is designed to extract the anatomical feature f_c from structure image x, i.e., $f_c = E_c(x), (f_c \in \mathbb{R}^{c'})$. Besides, E_c shares the same network architecture with E_p. It begins with a 3×3 convolutional layer, followed by 3 down-sampling ResBlocks to reduce the spatial dimensions and a standard ResBlock to refine the features. The output is then transformed into a vector through average pooling and passed through two linear layers with LeakyReLU activations.

Denoiser S tries to reconstruct the clean prompt vector, i.e., $\tilde{f}_{p,int}^0$, when given the anatomical feature f_c, timestep t, and noisy prompt vector f_p^t, i.e., $\tilde{f}_{p,int}^0 = S(f_c, t, f_p^t)$. Concretely, f_c, t, and f_p^t are input with concatenation and S employs a relatively simple network that only involves four identical Linear-LeakyReLU blocks.

Dual-level Reconstruction Constraints. Obtaining the predicted clean prompt vector $\tilde{f}_{p,int}^0$, we design dual-level (i.e., feature-level and image-level) reconstruction losses to collaboratively promote the reliability of prompts synthesized by the prompt generator. Concretely, given $\tilde{f}_{p,}^0$, we first use the following feature-level reconstruction loss L_{rec}^f to measure the disparity between the predicted prompt vector $\tilde{f}_{p,int}^0$ and its target f_p^0:

$$L_{rec}^f = \left\| f_p^0 - \tilde{f}_{p,int}^0 \right\|_1, \tag{11}$$

Furthermore, $\tilde{f}_{p,int}^0$ is fed into the pre-trained and parameter fixed prediction network P to reconstruct dose map $\hat{y}$ and we gain the following image-level reconstruction loss L_{rec}^i for better image-level reconstruction consistency:

$$L_{rec}^i = \|y - \hat{y}\|_1, \tag{12}$$

Comprehensively, we derive the entire loss of the prompt generation stage as below:

$$L = L_{rec}^i + \omega_1 L_{rec}^f + \omega_2 L_{adv}. \tag{13}$$

where ω_1 and ω_2 are two weighted hyperparameters to balance the three terms.

Table 1. Quantitative comparison results with SOTA methods on PTV. The best results are marked in **bold** while the second-best ones are underlined. *: $p < 0.05$ via the paired t-test.

Methods	ΔD_{95} (Gy) ↓	ΔD_2 (Gy) ↓	ΔD_{mean}(Gy) ↓	ΔCI ↓	Params (M)	iteration
U-ResNet-D	2.556±1.684*	1.043±0.383*	2.416±1.504*	0.087±0.127*	14.72	1
MTDP	1.769±1.372*	0.692±0.493*	1.305±1.116*	0.070±0.140*	13.86	1
PRUNet	1.531±1.955*	0.754±0.422*	0.959±1.465*	0.065±0.136*	25.42	1
Mc-GAN	1.757±1.514*	1.004±0.458*	1.268±1.049*	0.083±0.150*	1.15	1
TcTrans	1.599±1.565*	1.709±0.431*	1.427±1.029*	0.067±0.120*	94.07	1
SR3†	2.902±1.227*	0.817±0.507*	1.771±0.764*	0.167±0.131*	16.53	100
DiffDP	**1.195±1.613**	0.486±0.350*	0.780±1.180	0.068±0.131*	37.89	100
ViPDose	1.208±1.721	**0.292±0.219**	**0.732±1.238**	**0.062±0.128**	15.65	4

Table 2. Quantitative comparison results on OARs with regard to ΔD_{mean}(Gy).

	Sin	L.hf	R.hf	Bla
U-ResNet-D	2.250±1.634*	3.558±2.660*	3.109±2.130*	2.694±1.582*
MTDP	1.703±1.354*	2.383±1.826*	2.331±1.704*	2.118±1.545*
PRUNet	2.071±1.768*	2.747±2.095*	1.986±1.450*	2.528±1.987*
Mc-GAN	1.635±1.267	2.498±1.724*	2.128±1.836*	2.717±1.987*
TcTrans	1.539±1.481	2.272±1.994*	2.076±1.570*	2.074±1.624*
SR3†	**1.476±1.168**	1.977±1.811*	1.819±1.829*	1.399±1.050
DiffDP	2.786±1.954*	3.072±2.683*	2.431±1.940*	1.852±1.584*
ViPDose	1.498±1.422	**1.763±1.404**	**1.456±0.823**	**1.338±1.149**

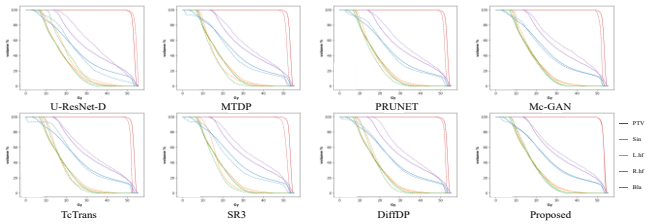

Fig. 2. The DVH curves for comparison with SOTA methods. Solid line represents the ground truth and dotted one represents the predicted dose maps.

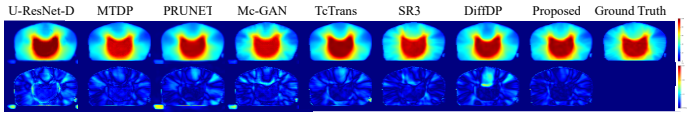

Fig. 3. Visual comparisons with SOTA methods. Top: dose maps, Bottom: error maps.

Table 3. Ablation results with different model variants on PTV and OARs.

Methods	PTV		R.hf	Bla
	ΔCI ↓	ΔD_{mean} (Gy) ↓	ΔD_{mean} (Gy) ↓	ΔD_{mean} (Gy) ↓
(A)	0.080±0.117	1.772±1.372	2.684±1.693	2.182±1.885
(B)	0.068±0.127	1.050±1.075	1.990±1.325	1.437±1.169
(B)‡	0.074±0.113	1.383±1.527	2.495±1.698	1.955±1.445
(C)	0.062±0.132	0.793±1.254	1.734±1.099	1.389±1.236
(D)	**0.062±0.128**	**0.732±1.238**	**1.456±0.823**	**1.338±1.149**

3 Experiments and Results

Dataset Descriptions and Evaluation Metrics. We employ an in-house clinical rectal cancer dataset with 130 patients to verify the superiority of the proposed ViPDose. All the patients have taken VMAT in West China Hospital. For each case, the CT scan, segmentation masks of PTV and OARs, and the clinically approved dose map are contained. The OARs involve small intestine (Sin), right head of the femur (R.hf), left head of the femur (L.hf), and bladder (Bla). We randomly choose 98/10/22 cases for training, validation and testing, respectively. 3D volumes with a resolution of 3mm×3mm×3mm are sliced along the axial direction, gaining 2D slices with a size of 160×160. Then, we introduce our evaluation metrics. For PTV, assigning Dx to denote the minimal absorbed dose which covers x% volume of PTV, we involve D_{95}, D_2, and mean dose (D_{mean}) as metrics. For OARs, we take D_{mean} for performance evaluation. Also, the conformation index (CI) [28] is used. We calculate the average mean error (Δ) for all quantitative metrics. Besides, the dose volume histogram (DVH) [29] is utilized.

Implementation Details. All experiments are conducted on a single NVIDIA GeForce RTX 3090 GPU equipped with 24GB of memory. The two phases of ViPdose are both trained with 200 epochs with a batch size of 16. The Adam optimizer is utilized to promote efficient convergence. The learning rates of the two phases are set as 5e-6 and 3e-6, respectively. The hyper-parameters ω_1 and ω_2 in Eq. (13) are empirically set as 1 and 0.1, respectively. The iteration T for the DAN is 4 for both training and testing

Comparison Studies. To validate the performance of ViPDose, we comprehensively compare it with non-diffusion and diffusion-based methods. Non-diffusion methods comprise U-ResNet-D [11], MTDP [10], PRUNet [12], Mc-GAN [17], and TcTrans [9]; diffusion-based methods involve SR3 [23] and DiffDP [18]. Notably, SR3 is a classical DM for super resolution which serves as a baseline of image-conditional DM. Here, we adopt it to the dose prediction task for further comparison (marked by "†"). The quantitative results on PTV and OARs are reported in Table 1 and Table 2, respectively. In Table 1, compared to non-diffusion methods, ViPDose gains notably better performance, especially obtaining 1.208Gy in terms of ΔD_{95}. Moreover, compared to the diffusion-based method DiffDP, ViPDose reaches higher accuracy for the rest three metrics. SR3 obtains the second-best accuracy on four OARs but our ViPDose still maintains its leading performance. The DVH curve in Fig.2 shows that our proposed ViPDose gains the smallest disparities to the ground truth for all organs which indicates its superiority. Visual comparisons with SOTA methods are given in Fig. 3 where the proposed method generates more realistic dose maps with the minimal error maps. Table 1 also shows the relatively less parameters and iterations of the proposed method.

Ablation Studies. The arrangements of ablation models are: (A) Prediction network without prompt as Baseline, (B) Baseline + Prompt generated by DM trained with feature-level reconstruction loss L_{rec}^{f} only (Baseline + Prompt), (C) Baseline + Prompt +

L_{adv}, and (D) Baseline + Prompt + L_{adv} + L_{rec}^{i} (Proposed). As seen in Table 3, after gradually adding proposed components, the performance is progressively enhanced, verifying their respective contributions. Furthermore, we design an exploratory experiment to additionally investigate whether DM has the superiority in generating prompt vectors compared to regression-based model. We train the denoiser to directly generate $\tilde{f}_{p,t}^{0}$ with only L_{rec}^{f} and denoted it as (B)‡. Compared to (B)‡, (B) gains higher accuracy.

4 Conclusion

In this work, we present ViPDose to utilize the visual prompt of dose maps to effectively achieve automatic dose distribution in radiotherapy. We train the prompt encoder to compress the dose maps into prompt vectors in the prompt pre-training stage. Then, we use a DAN to predict the prompt vectors during the prompt generation stage, thus fully exploring the dose distribution knowledge. Experiments on the clinical rectal cancer dataset have sufficiently verified its leading performance.

Acknowledgments. This work is supported by National Natural Science Foundation of China (NSFC 62371325, U23A20295, 62131015, 82441023, 82394432), Sichuan Science and Technology Program (2025NSFJQ0050, 2024ZDZX0018), and Key Lab of Internet Natural Language Processing of Sichuan Provincial Education Department (No.INLP202402).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Thariat, J., Hannoun-Levi, J. M., Sun Myint, A., Vuong, T., Gérard, J. P.: Past, present, and future of radiotherapy for the benefit of patients. Nature reviews Clinical oncology, 10(1), 52-60 (2013).
2. Murakami, Y., Nakano, M., Yoshida, M., Hirashima, H., Nakamura, F., Fukunaga, J., and Hirata, H.: Possibility of chest wall dose reduction using volumetric-modulated arc therapy (VMAT) in radiation-induced rib fracture cases: comparison with stereotactic body radiation therapy (SBRT). Journal of Radiation Research, 59(3), 327-332 (2018).
3. Gao, R., Lou, B., Xu, Z., Comaniciu, D., and Kamen, A. Flexible-Cm GAN: Towards Precise 3D Dose Prediction in Radiotherapy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 715-725 (2023).
4. Wang, B., Teng, L., Mei, L., Cui, Z., Xu, X., Feng, Q., and Shen, D.: Deep Learning-Based Head and Neck Radiotherapy Planning Dose Prediction via Beam-Wise Dose Decomposition. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 575-584 (2022).
5. Cui, J., Zeng, P., Zeng, X., Xu, Y., Wang, P., Zhou, J., Wang, Y., and Shen, D.: Prior Knowledge-Guided Triple-Domain Transformer-GAN for Direct PET Reconstruction From Low-Count Sinograms. IEEE Transactions on Medical Imaging 43(12), 4174–4189 (2024).
6. Wang, Y., Zhou, L., Yu, B., Wang, L., Zu, C., Lalush, D.S., Lin, W., Wu, X., Zhou, J., and Shen, D.: 3D Auto-Context-Based Locality Adaptive Multi-Modality GANs for PET Synthesis. IEEE Transactions on Medical Imaging 38(6), 1328–1339 (2019).

7. Wang, K., Zhan, B., Zu, C., Wu, X., Zhou, J., Zhou, L., and Wang, Y.: Semi-supervised medical image segmentation via a tripled-uncertainty guided mean teacher model with contrastive learning. Medical Image Analysis 79, 102447 (2022).
8. Cui, J., Wang, Y., Zhou, L., Fei, Y., Zhou, J., Shen, D.: 3D Point-Based Multi-Modal Context Clusters GAN for Low-Dose PET Image Denoising. IEEE Transactions on Circuits and Systems for Video Technology 34(10), 9400–9413 (2024).
9. Jiao, Z., Peng, X., Wang, Y., Xiao, J., Nie, D., Wu, X., Wang, X., Zhou, J., and Shen, D.: TransDose: Transformer-based radiotherapy dose prediction from CT images guided by super-pixel-level GCN classification. Medical Image Analysis 89, 102902 (2023).
10. Tan, S., Tang, P., Peng, X., Xiao, J., Zu, C., Wu, X., and Wang, Y.: Incorporating isodose lines and gradient information via multi-task learning for dose prediction in radiotherapy. In Medical Image Computing and Computer Assisted Intervention, Proceedings, Part VII 24, pp. 753-763 (2021).
11. Liu, Z., Fan, J., Li, M., Yan, H., Hu, Z., Huang, P., and Dai, J.: A deep learning method for prediction of three-dimensional dose distribution of helical tomotherapy. Medical physics, 46(5), 1972-1983 (2019).
12. Wang, J., Hu, J., Song, Y., Wang, Q., Zhang, X., Bai, S., and Yi, Z.: VMAT dose prediction in radiotherapy by using progressive refinement UNet. Neurocomputing, 488, pp. 528-539 (2022).
13. Wen, L., Zhang, Q., Feng, Z., Xu, Y., Chen, X., Zhou, J., and Wang, Y.: Triplet-constraint Transformer with Multi-scale Refinement for Dose Prediction in Radiotherapy. In 2024 IEEE 21th International Symposium on Biomedical Imaging (ISBI), IEEE (2024).
14. Liao, M., Di, S., Zhao, Y., Liang, W., and Yang, Z.: FA-Net: A hierarchical feature fusion and interactive attention-based network for dose prediction in liver cancer patients. Artificial Intelligence in Medicine 156, 102961 (2024).
15. Gheshlaghi, T., Nabavi, S., Shirzadikia, S., Moghaddam, M.E., and Rostampour, N.: A cascade transformer-based model for 3D dose distribution prediction in head and neck cancer radiotherapy. Physics in Medicine & Biology 69(4), 045010 (2024).
16. Mahmood, R., Babier, A., McNiven, A., Diamant, A., and Chan, T. C.: Automated treatment planning in radiation therapy using generative adversarial networks. In Machine learning for healthcare conference, pp. 484-499 (2018).
17. Zhan, B., Xiao, J., Cao, C., Peng, X., Zu, C., Zhou, J., and Wang, Y.: Multi-constraint generative adversarial network for dose prediction in radiotherapy. Medical Image Analysis, 77, 102339 (2022).
18. Feng, Z., Wen, L., Wang, P., Yan, B., Wu, X., Zhou, J., and Wang, Y.: DiffDP: Radiotherapy dose prediction via a diffusion model. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 191-201 (2023).
19. Cui, J., Jiao, Z., Wei, Z., Hu, X., Wang, Y., Xiao, J., and Peng, X.: CT-only radiotherapy: An exploratory study for automatic dose prediction on rectal cancer patients via deep adversarial network. Frontiers in Oncology 12, 875661 (2022).
20. Ronneberger, O., Fischer, P., and Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention, Part III 18, pp. 234-241 (2015).
21. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., and Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems, 27 (2014).
22. Ho, J., Jain, A., and Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems, 33, 6840-6851 (2020).

23. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., and Norouzi, M.: Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4), 4713-4726 (2022).
24. Kim, B., and Ye, J. C.: Diffusion deformable model for 4D temporal medical image generation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 539-548 (2023).
25. He, X., Tan, C., Han, L., Liu, B., Axel, L., Li, K., and Metaxas, D. N.: DMCVR: Morphology-Guided Diffusion Model for 3D Cardiac Volume Reconstruction. In International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 132-142 (2023).
26. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B.: High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684-10695 (2022).
27. Xiao, Z., Kreis, K., and Vahdat, A.: Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. In International Conference on Learning Representations.
28. Van't Riet, A., Mak, A. C., Moerland, M. A., Elders, L. H., and Van Der Zee, W.: A conformation number to quantify the degree of conformality in brachytherapy and external beam irradiation: application to the prostate. International Journal of Radiation Oncology* Biology* Physics, 37(3), 731-736 (1997).
29. Graham, M. V., Purdy, J. A., Emami, B., Harms, W., Bosch, W., Lockett, M. A., and Perez, C. A.: Clinical dose–volume histogram analysis for pneumonitis after 3D treatment for non-small cell lung cancer (NSCLC). International Journal of Radiation Oncology* Biology* Physics, 45(2), 323-329 (1999).