Automated Radiation Therapy Treatment Planning using 3-D Generative Adversarial Networks

Aaron Babier, Rafid Mahmood[*]
Mechanical & Industrial Engineering
University of Toronto
{ababier,rmahmood}@mie.utoronto.ca

Andrea McNiven
Radiation Medicine Program
Princess Margaret Cancer Centre
andrea.mcniven@rmp.uhn.ca

Adam Diamant
Schulich School of Business
York University
adiamant@schulich.yorku.ca

Timothy C. Y. Chan
Mechanical & Industrial Engineering
University of Toronto
tcychan@mie.utoronto.ca

Abstract

Knowledge-based automated planning (KBAP) is a data-driven radiation therapy treatment planning method that first predicts desirable treatments before generating deliverable ones. We propose the first 3-D generative adversarial network-based KBAP pipeline that predicts a dose distribution from a CT scan before optimizing for deliverability. Our experiments on a dataset of oropharyngeal cancer patients show that this new framework, 3-D GANCER, significantly outperforms previous methods on replicating the same clinical criteria satisfaction as real plans.

1 Introduction

Radiation therapy (RT) is a common cancer treatment method and prescribed for nearly 50% of all cases [Delaney et al., 2005]. An RT treatment plan has two components: the dose distribution tensor whose elements are the dose delivered to each voxel of the patient's body, and the vector of beamlet intensities from a linear accelerator (LINAC) that delivers the radiation. Designing a treatment plan is a complex problem where beamlet intensities are optimized to satisfy clinical constraints, while minimizing multiple competing objectives. The current clinical practice is an iterative back-and-forth between a dosimetrist who generates a treatment plan, and an oncologist who suggests revisions. Completing a single plan may take several days, making the procedure both labor intensive and costly.

Contrasting the iterative procedure, Knowledge-Based Automated Planning (KBAP) is a data-driven approach that trains on historical plans to generate new plans for future patients. KBAP contains two stages: (a) a machine learning model that predicts a clinically satisfactory dose [Appenzoller et al., 2012, Younge et al., 2018], and (b) an optimization model that prescribes a deliverable treatment plan (i.e., beamlets and dose) [McIntosh and Purdie, 2017, Babier et al., 2018a]. Prediction is essential to KBAP, as the quality of final plans strongly correlate with the quality of predicted doses [Babier et al., 2018a]. Thus, the literature focuses on designing better predictive models. Historically, these have been classical ML models that predict low-dimensional representations of the dose (e.g., summary statistics and histograms) [Zhu et al., 2011, Yuan et al., 2012, Babier et al., 2018a]. Only recently have researchers introduced models to directly predict the full dose distribution, with their results significantly outperforming classical methods [McIntosh and Purdie, 2016, Mahmood et al., 2018].

Mathematically, the dose delivered is simply a 3-dimensional heatmap. Therefore, predicting dose is a matter of inputting the 3-D CT image of a patient and outputting a 3-D tensor representing dose.

[*] A. Babier and R. Mahmood contributed equally to this work.

Machine Learning for Health (ML4H) Workshop at NeurIPS 2018.

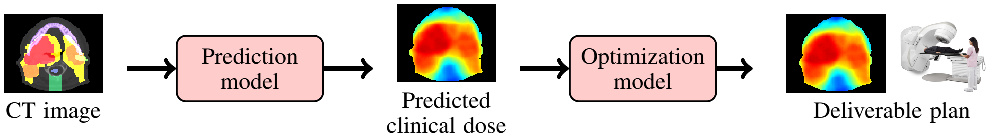


Figure 1: Overview of the KBAP pipeline.

This is a style transfer problem and it is well understood that generative adversarial networks (GANs) excel in this task [Goodfellow et al., 2014, Wu et al., 2016, Isola et al., 2017]. Moreover, GANs have been studied in limited extent for KBAP [Mahmood et al., 2018]. This work aims to develop a complete study of GANs as the prediction engine in KBAP. Specifically, we contribute the following:

1. We design the first 3-D style transfer GAN for KBAP that inputs a patient's CT image and outputs the dose distribution. This is the first neural network-based approach that predicts the entire 3-D dose tensor without feature engineering from the original 3-D CT image.
2. We suggest that representing dose intensity as gray-scale (i.e., a 1-channel output) produces better predictions than if the full 3-channel (RGB) input is used.
3. We compare our model, referred to as 3-D GANCER, with the state-of-the-art on a dataset of 217 oropharyngeal cancer patients. We implement the full KBAP pipeline and evaluate the post-optimization plans. Our model outperforms all baselines on several clinical metrics.

2 Methods

Fig. 1 highlights the KBAP pipeline. Patient CT scans are used to predict (via our model or baselines) a clinically acceptable dose distribution. Predictions are used in an optimization model to generate a plan [Babier et al., 2018b]. A review of KBAP is provided in McIntosh and Purdie [2016].

2.1 Data

We obtained plans for 217 oropharyngeal cancer patients treated at a single institution with 6 MV, step-and-shoot, intensity-modulated radiation therapy. All plans had at least two planning target volumes (PTVs) that were prescribed 70 Gy and 56 Gy in 35 fractions to the gross disease (PTV70) and elective target volumes (PTV56), respectively; 130 plans also had a prescription of 63 Gy to an intermediate risk target volume (PTV63). The brainstem, spinal cord, right parotid, left parotid, larynx, esophagus, mandible, and the limPostNeck (an artificial structure in the posterior neck) are referred to as organs-at-risk (OAR), and must be spared sufficiently according to clinical criteria.

Each patient is discretized into voxels of size 4 mm $\times$ 4 mm $\times$ 2 mm. CT images and dose distributions are converted into $128 \times 128 \times 128$ tensors of 3-channel (RGB) and 1-channel (dose intensity) respectively. We randomly spit the data into 130 patients for training and 87 for testing.

2.2 Generative Adversarial Network

In KBAP, a style transfer GAN consists of two networks that learn to take a CT image $\mathbf{c}$ and random initialization $\mathbf{z} \sim p_{\mathbf{z}}$ and predict the dose $\mathbf{x} \sim p_{\text{data}}$ [Goodfellow et al., 2014, Isola et al., 2017]. Specifically, a generator network $G(\mathbf{c}, \mathbf{z}) = \mathbf{x}$ performs the prediction while a discriminator $D(\mathbf{x}, \mathbf{c}) \in [0, 1]$ learns to classify generator output. We use the conventional style transfer loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}} \left[\log D(\mathbf{x}, \mathbf{c})\right] + \mathbb{E}_{\mathbf{z} \sim p_{\mathbf{z}}} \left[\log(1 - D(G(\mathbf{z}, \mathbf{c}), \mathbf{c}))\right] + \lambda \mathbb{E}_{\mathbf{x} \sim p_{\text{data}}, \mathbf{z} \sim p_{\mathbf{z}}} \left[\|\mathbf{x} - G(\mathbf{z}, \mathbf{c})\|\right].$$

The generator minimizes the last two terms of this loss function, while the discriminator maximizes the first two terms. While the first two terms comprise the standard GAN loss function, the third is used in style transfer problems to help generate images that better resemble the ground truth. The hyperparameter λ balances the tradeoffs between the standard GAN loss and the style transfer.

The generator and discriminator network architectures were derived from the U-net `pix2pix` of Isola et al. [2017], with the key difference being the use of 3-D convolutional filters. We provide more details on the network architecture, training, and parameter selection in Appendix A.

Table 1: The percentage of final plans of each KBAP population that satisfy the same clinical criteria as the clinical plans. Here, $\mathcal{D}_{mean}$ is the mean dose to the structure, $\mathcal{D}_{max}$ is the maximum dose, and $\mathcal{D}_{99}$ is the dose to the 99-th percentile. Highest percentages per structure are in bold.

Structure	Criteria	RF	2-D RGB-GAN	2-D GANCER	3-D GANCER
Brain Stem	$\mathcal{D}_{max} \leq 54$ Gy	**100.0**	**100.0**	**100.0**	**100.0**
Spinal Cord	$\mathcal{D}_{max} \leq 48$ Gy	**100.0**	98.8	98.9	**100.0**
Right Parotid	$\mathcal{D}_{mean} \leq 26$ Gy	95.2	**98.8**	97.6	**98.8**
Left Parotid	$\mathcal{D}_{mean} \leq 26$ Gy	94.2	94.2	**97.7**	**97.7**
Larynx	$\mathcal{D}_{mean} \leq 45$ Gy	92.5	92.4	92.5	**95.0**
Esophagus	$\mathcal{D}_{mean} \leq 45$ Gy	**100.0**	**100.0**	**100.0**	**100.0**
Mandible	$\mathcal{D}_{max} \leq 73.5$ Gy	80.5	73.6	**88.5**	86.2
PTV70	$\mathcal{D}_{99} \geq 66.5$ Gy	98.9	83.9	**100.0**	**100.0**
PTV63	$\mathcal{D}_{99} \geq 59.9$ Gy	98.0	**100.0**	**100.0**	**100.0**
PTV56	$\mathcal{D}_{99} \geq 53.2$ Gy	97.7	94.3	97.7	**100.0**
All structures		66.7	48.3	75.9	**78.2**

2.3 Plan optimization

Predicted dose distributions are input into an inverse optimization (IO) pipeline [Babier et al., 2018b]. The IO model uses the predicted dose distribution to determine parameters of a treatment planning optimization model, which is then solved to generate the plan. The planning model minimizes a weighted sum of 65 objective functions: seven per OAR and three per PTV. The OAR objectives include the mean dose, maximum dose, and the average dose above 0.25, 0.50, 0.75, 0.90, and 0.975 of the maximum predicted dose to that OAR. Target objectives include the maximum dose, average dose below prescription, and average dose above prescription. In addition, the sum-of-positive-gradients (SPG), which is a surrogate measure for the physical deliverability from a LINAC, is constrained to be less than 55 [Craft et al., 2007]. The influence of beamlets to dose voxels are derived using `CERR` [Deasy et al., 2003]. We used `Gurobi 7.5` to solve the optimization problem.

2.4 Baselines

We compare our model with several state-of-the-art baselines. All predictions are passed through the IO model in the same way as our GAN predictions. Training details are given in Appendix A.

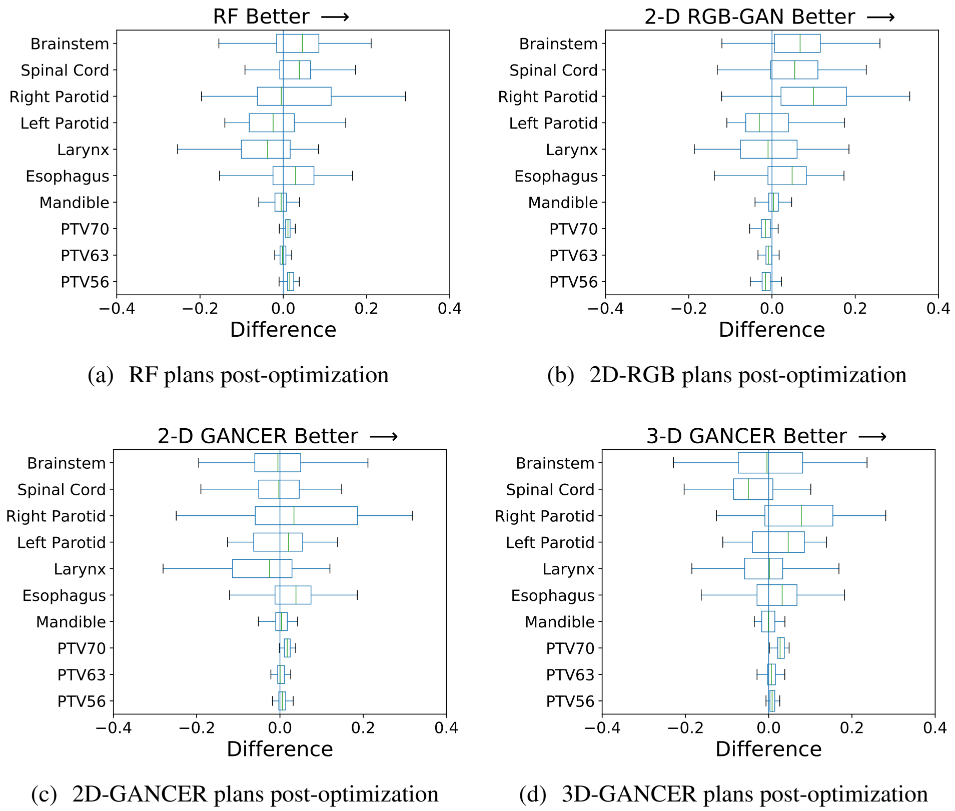
- **2-D RGB-GAN:** Predicts 128 RGB images representing horizontal slices of the dose distribution using the `pix2pix` architecture (e.g., [Mahmood et al., 2018]).
- **2-D GANCER:** 1-channel variant of 2-D RGB-GAN that predicts 128 slices of the dose.
- **Random Forest (RF):** Predicts dose to each voxel using 133 features and Gaussian filters (e.g., McIntosh and Purdie [2017]). The state-of-the-art from classical ML models.

3 Results

A sample of the prediction and post-prediction manifold is available in Appendix B. We evaluated the models using 10 clinical criteria (Table 1). We counted the frequency of plans that satisfied the same criteria as the ground truth clinical plans. We also calculated the extent to which each criteria was satisfied. Clinical motivation and additional details on the metrics are provided in Appendix B.

Table 1 lists the frequency of KBAP plans that satisfy the same clinical criteria as the corresponding clinical plans. For example, 95% of the fraction of clinical plans that satisfy the larynx criteria, have corresponding 3-D GANCER plans that also satisfy this criteria. Overall, we find that 3-D GANCER plans consistently replicate clinical plan quality for all PTVs and a majority of OARs. Moreover, 3-D GANCER plans consistently outperform the baselines on all criteria, except for the mandible. This is likely a consequence of good performance on targets, which are geometrically close to the mandible.

Figure 2 shows the distribution of differences among criteria between KBAP versus corresponding clinical plans. 2-D and 3-D GANCER plans are most similar to clinical, and on average perform 0.3% better on criteria, but 3-D GANCER plans have a lower average inter-quartile range (IQR) (0.08 vs 0.09). While 2-D RGB-GAN best outperforms clinical plans on sparing criteria (average



(a) RF plans post-optimization

(b) 2D-RGB plans post-optimization

(c) 2D-GANCER plans post-optimization

(d) 3D-GANCER plans post-optimization

Figure 2: The distribution of criteria differences between the KBAP plans and the clinical plans.

2.4% improvement, 0.08 IQR), they fail to meet PTV criteria, which is unacceptable. RF plans are on average 1.1% better than clinical with the smallest IQR (0.08), but Table 1 shows they do not always satisfy the same criteria as clinical. Predicting the dose directly using GANs yields plans that are most similar to clinical; in addition, 3-D GANCER plans are the most consistent in the lower IQR.

4 Discussion and Future Work

We propose the first 3-D style transfer GAN for KBAP, which learns to predict the entire dose distribution from a CT scan. We implement our model on a oropharyngeal cancer dataset along with several baselines. We compare all models by evaluating post-optimization final plans and find that our model, 3-D GANCER, outperforms all baselines on standard clinical criteria for evaluating RT treatment plans. Compared to prior work on 2-D slices, a 3-D style transfer GAN improves predictions by accounting for correlations in the vertical axis. For example, a real dose distribution will decrease smoothly beyond the boundary of the PTV; a 2-D style transfer GAN is likely to predict the dose dropping sharply along this edge. Permitting vertical correlations allows us to consistently replicate criteria satisfaction of historical, oncologist-approved plans for nearly every patient organ.

While extremely powerful, GANs have historically been hard to train, i.e., outputting hallucinations, or mode collapse. However, we find that GANs perform without difficulty inside a KBAP pipeline, often with just off-the-shelf architectures and default parameters. We believe this success is due to two factors. First, in contrast to a conventional task of generating detailed photographs, the task of predicting dose tensors is much easier, as these are smooth images with relatively simple characteristics. Second, all predictions are passed through an optimization model to obtain "physically deliverable" plans. Thus, not only do GANs face an easier task than in conventional applications, but the optimization model also acts as a safety-net that corrects potential hallucinations or errors. In KBAP, the weaknesses of GANs are nullified, making them very appropriate for dose prediction.

Our results suggest 3-D GANCER is the most effective predictive component to date in KBAP for oropharyngeal cancer. Future work should extend this method to other cancer sites.

Acknowledgments

This study was approved by the institutional research ethics board. Support for this research was provided by the Natural Sciences and Engineering Research Council of Canada

References

L. M. Appenzoller, J. M. Michalski, W. L. Thorstad, S. Mutic, and K. L. Moore. Predicting dose-volume histograms for organs-at-risk in imrt planning. Medical physics, 39(12):7446–7461, 2012.

A. Babier, J. J. Boutilier, McNiven A. L., M. B. Sharpe, and T. C. Y. Chan. Knowledge-based automated planning for oropharyngeal cancer. Med Phys, 45(7):2875–2883, Jul 2018a. doi: 10.1002/mp.12930.

A. Babier, J. J. Boutilier, M. B. Sharpe, A. L. McNiven, and T. C. Y. Chan. Inverse optimization of objective function weights for treatment planning using clinical dose-volume histograms. Phys Med Biol, 63(10): 105004, May 2018b. doi: 10.1088/1361-6560/aabd14.

D. Craft, P. Suss, and T. Bortfeld. The tradeoff between treatment plan quality and required number of monitor units in intensity-modulated radiotherapy. Int. J. Radiat. Oncol. Biol. Phys., 67(5):1596–605, 2007.

J. O. Deasy, A. I. Blanco, and V. H. Clark. CERR: a computational environment for radiotherapy research. Med. Phys., 30(5):979–85, 2003.

G. Delaney, S. Jacob, C. Featherstone, and M. Barton. The role of radiotherapy in cancer treatment. Cancer, 104 (6):1129–1137, 2005.

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In Advances in neural information processing systems, pages 2672–2680, 2014.

P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. arXiv preprint, 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

R. Mahmood, A. Babier, A. Mcniven, A. Diamant, and T.C.Y. Chan. Automated treatment planning in radiation therapy using generative adversarial networks. Proceedings of Machine Learning Research, 2018.

C. McIntosh and T. G. Purdie. Contextual atlas regression forests: Multiple-atlas-based automated dose prediction in radiation therapy. IEEE Trans Med Imaging, 35(4):1000–12, Apr 2016. doi: 10.1109/TMI.2015.2505188.

C. McIntosh and T. G. Purdie. Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. Phys Med Biol, 62(2):415–431, Jan 2017. doi: 10.1088/1361-6560/62/2/415.

J. Wu, C. Zhang, T. Xue, B. Freeman, and J. Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In Advances in Neural Information Processing Systems, pages 82–90, 2016.

K. C. Younge, R. B. Marsh, D. Owen, H. Geng, Y. Xiao, D. E. Spratt, J. Foy, K. Suresh, Q. J. Wu, F. Yin, S. Ryu, and M. M. Matuszak. Improving quality and consistency in nrg oncology radiation therapy oncology group 0631 for spine radiosurgery via knowledge-based planning. Int J Radiat Oncol Biol Phys, 100(4):1067–1074, Mar 2018. doi: 10.1016/j.ijrobp.2017.12.276.

L. Yuan, Y. Ge, W. R. Lee, F. F. Yin, J. P. Kirkpatrick, and Q. J. Wu. Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans. Med. Phys., 39(11):6868–78, 2012.

X. Zhu, Y. Ge, T. Li, D. Thongphiew, F. Yin, and Q. J. Wu. A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. Med. Phys., 38(2):719–26, 2011.

A Model details

A.1 Training the GANs

For all of the GANs used in experiments, both the generator and discriminator networks are derived from the `pix2pix` architecture [Isola et al., 2017], We use a U-net architecture for the generator that passes contoured CT images through consecutive convolution layers, a bottleneck layer, and then through several deconvolution layers. The U-net also employs skip connections, i.e., the output of each convolution layer is concatenated to the input of a corresponding deconvolution layer. This allows the generator to easily pass "high-dimensional" information (e.g., structural outlines) between the inputted CT image and the outputted dose. The discriminator passes a dose distribution along several consecutive convolution layers, outputting a single scalar value. We use the same number of layers in the encoder, bottleneck, and decoder. The 2-D RGB-GAN architecture, which was proposed in Mahmood et al. [2018] is an exactly the same (i.e., the same filter sizes) as the `pix2pix` in Isola et al. [2017]. The 2-D GANCER architecture is equivalent to Isola et al. [2017] except with the final layer outputting a 1 channel 2-D slice. The 3-D GANCER architecture is the same as the 2-D GANCER, except all filters are 3-D convolutional with the same sizes as before.

We used the loss function described in Section 2.2 with $\lambda = 90$, and trained using the Adam optimization algorithm [Kingma and Ba, 2014], with learning rate 0.0002 and $\beta_1 = 0.5$ and $\beta_2 = 0.999$. These hyperparameters are the default Adam settings [Isola et al., 2017], and are good for a variety of different style transfer problems. While we swept through various values for λ, we found the default to be sufficient. We also found it useful to stop training when the loss functions were roughly equal; if the loss from the l_1 penalty fell too low, the GAN began to memorize the dataset. To prevent this the 2-D RGB-GAN and 2-D GANCER networks were trained for 50 epochs, and the 3-D GANCER network was trained for 200 epochs. These settings are listed in Table 2.

In our initial experiments, we found that the final plans outputted from KBAP using 2-D GANCER and 3-D GANCER were actually underperforming by not providing sufficient dose to the PTV. Therefore, we rescaled the dose predictions of 2-D GANCER and 3-D GANCER by multiplying by a scalar until all of the target criteria were "just" satisfied. This scalar varies for each plan, but it is easy to determine with a simple sweep. Note that this scaling does not affect the fairness of the analysis as all predictions are regardless input into the IO model in order to obtain a "physically deliverable" treatment plan. That is, the final plans all have the same complexity regardless of any scaling post-prediction. Note that we did not rescale the 2-D RGB-GAN as we found that this scaling factor negatively impacted the performance of this model.

Table 2: GAN settings.

	2-D RGB-GAN	2-D GANCER	3-D GANCER
Generator input	Axial slice of CT	Axial slice of CT	Whole CT image
Generator output	128 × 128 RGB image	128 × 128 dose	128 × 128 × 128 dose
Post-prediction scaling	No	Yes	Yes
Number of epochs	50	50	200
λ	90		
Learning rate	0.0002		
Adam momentum	$(\beta_1, \beta_2) = (1.05e-5, 0.999)$		

A.2 Training the Random Forest

The random forest used 133 features outlined in Table A.2 to predict the dose delivered to each voxel in the patient. Of these features, 10 of them were positional and hand-tailored, while 122 were generated by applying 122 Gaussian filters (GFs) to the grayscale CT image. One of the GFs was isotropic ($\sigma = 10$) and a second was the Laplacian of the Gaussian ($\sigma = 10$). The remaining 120 filters were made of all combinations of the following four parameters: (a) first and second order GFs, (b) $\sigma = 4, 12, 24, 48$, and 64, (c) rotations of $0, 90, 180$, and 270 degrees, and (d) rotations in each of the three axes. The RF was trained with 10 trees and default settings using the `randomForestRegressor` from `scikit-learn`.

B Analysis

Figure 3 shows a set of 5 slices of an average 3-D GANCER predicted dose distribution, comparing with the corresponding CT slice, clinical dose, and the dose corresponding to the final plan. 3-D GANCER appropriately learns to predict dose distributions that have the hallmarks of deliverability, i.e., sharp gradients generated from individual beams. However, there are some subtle characteristics that 3-D GANCER cannot always identify. Using IO post-prediction allows us to correct for these idiosyncrasies quite easily.

Table 3: The ten features used in the RF to predict the dose for any voxel.

Feature	Description
Structure	Structure that the voxel is classified as
x-coordinate	Voxel's positions on the x-axis in a slice
y-coordinate	Voxel's positions on the y-axis in a slice
z-coordinate	Plane of voxel's slice
Distance to larynx	Shortest path between voxel and the surface of the larynx
Distance to esophagus	Shortest path between voxel and the surface of the esophagus
Distance to limPostNeck	Shortest path between voxel the surface of the limPostNeck
Distance to PTV56	Shortest path between voxel and the surface of the PTV56
Distance to PTV63	Shortest path between voxel and the the surface of PTV63
Distance to PTV70	Shortest path between voxel and the the surface of PTV70
CT gray-scale	gray-scale of the voxel in the CT image
122 GF CT gray-scales	gray-scale of the voxel in the CT image post GF

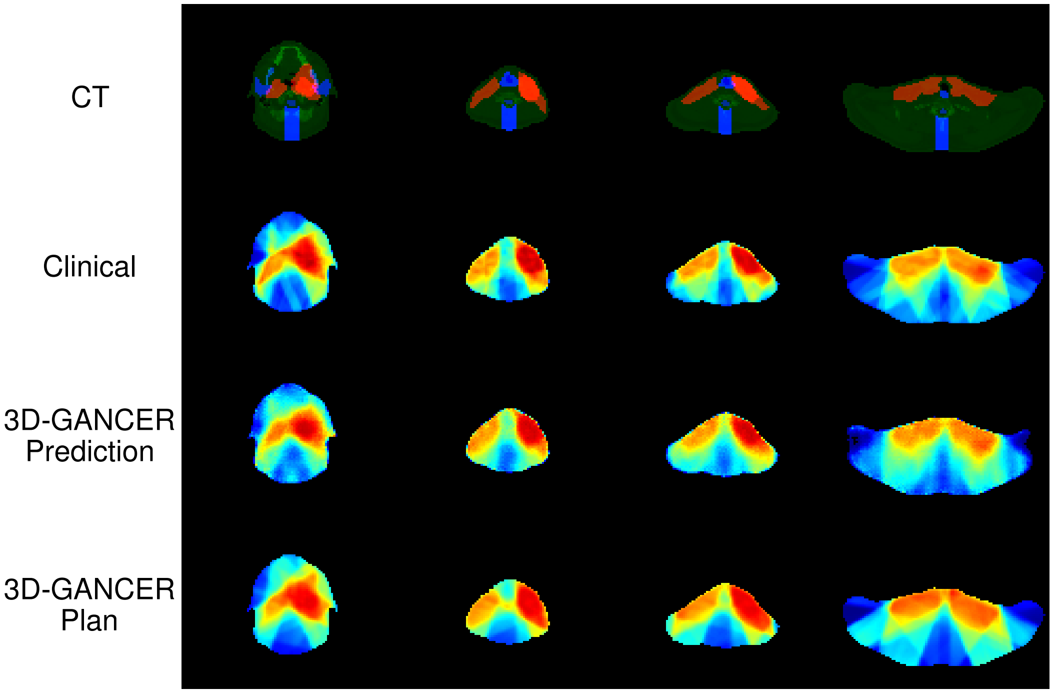


Figure 3: Sample of slices from a test patient. From top to bottom: contoured CT image (generator input), clinical plan (ground truth), 3D-GANCER prediction, and 3D-GANCER plan (post optimization). Note that the gray-scale dose distributions are represented as RGB for clarity.

In clinical practice, all treatment plans must be approved by an oncologist before delivered. The oncologist naturally evaluates several quantitative criteria along with some intuition in order to judge plan quality. Because physical studies with oncologists can be expensive, the standard procedure for evaluating automated treatment planning methods is to test if the plans successfully satisfy several quantitative criteria.

We use the criteria listed in Table 1 to evaluate our plans. These exact criteria have been used as benchmarks in a number of previous studies on KBAP for treating oropharyngeal cancer [Babier et al., 2018a,b, Mahmood et al., 2018]. There are 10 criteria, one for each OAR and PTV. The OAR criteria are generally testing whether the average or the maximum (depending on the specific structure) dose to all voxels in that structure is below the required value. The PTV criteria test whether the 99-th percentile dose to the specific PTV is at least greater than or equal to the oncologist prescribed dose.

B.1 Clinical Criteria Satisfaction

It is generally quite difficult to satisfy the clinical criteria, and in fact, the clinical plans in our dataset only satisfy 73% of the criteria on average. Therefore, the evaluation in Table 1 is counting the frequency of generated plans that satisfy the same criteria as the ground truth. For example, 95% of all patients whose clinical plans satisfied the larynx criterion have corresponding 3-D GANCER generated plans that also satisfy the larynx criterion. This does not preclude patients whose clinical plans did not satisfy the larynx criterion, but for whom the generated plans may have. This metric ensures that generated plans are as representative of the clinical as possible.

B.2 Clinical Criteria Performance

While clinical criteria satisfaction simply calculates whether the dose meets the threshold in the criteria, the clinical criteria performance metric evaluates the relative difference between the KBAP plan and the corresponding clinical plan on the dose achieved. For example, even if the clinical and KBAP plans satisfy the larynx criteria and achieve $\mathcal{D}_{mean} \leq 45$ Gy, the plan with the lower $\mathcal{D}_{mean}$ is sparing the larynx better than the other. Furthermore, by breaking down the differences on each constraint, we can better understand the tradeoffs made by each KBAP pipeline. For example, the post-optimization plans generated via 2-D RGB-GAN on average significantly outperform the clinical plans on criteria for sparing healthy tissue. In fact, they appear to outperform nearly every other method for the brainstem, spinal cord, and right parotid. However, this comes at a cost of not successfully meeting the PTV criteria, which are arguably more important. In contrast, the 3-D GANCER plans are most similar to clinical plans, but only slightly outperforming them on most criteria. Ideally, an automated planning method would generate plans that pass the same criteria as their clinical counterparts, but amplify the margins on those specific criteria.