Learning Treatment Plan Representations for Content Based Image Retrieval

Charles Huang 1, Varun Vasudevan 3, Oscar Pastor-Serrano 2,4, Md Tauhidul Islam 2, Yusuke Nomura 2, Yong Yang 2 and Lei Xing 2

1 Department of Bioengineering, Stanford University, Stanford, USA
2 Department of Radiation Oncology, Stanford University, Stanford, USA
3 Institute for Computational & Mathematical Engineering, Stanford University, Stanford, USA
4 Department of Radiation Science and Technology, Delft University of Technology, the Netherlands

E-mail: xxx@xxx.xx

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Objective: Knowledge based planning (KBP) typically involves training an end-to-end deep learning model to predict dose distributions. However, training end-to-end KBP methods may be associated with practical limitations due to the limited size of medical datasets that are often used. To address these limitations, we propose a content based image retrieval (CBIR) method for retrieving dose distributions of previously planned patients based on anatomical similarity.

Approach: Our proposed CBIR method trains a representation model that produces latent space embeddings of a patient’s anatomical information. The latent space embeddings of new patients are then compared against those of previous patients in a database for image retrieval of dose distributions. Summary metrics (e.g. dose-volume histogram, conformity index, homogeneity index, etc.) are computed and can then be utilized in subsequent automated planning. All source code for this project is available on github.

Main Results: The retrieval performance of various CBIR methods is evaluated on a dataset consisting of both publicly available plans and clinical plans from our institution. This study compares various encoding methods, ranging from simple autoencoders to more recent Siamese networks like SimSiam, and the best performance was observed for the multitask Siamese network.

Significance: Applying CBIR to inform subsequent treatment planning potentially addresses many limitations associated with end-to-end KBP. Our current results demonstrate that excellent image retrieval performance can be obtained through slight changes to previously developed Siamese networks. We hope to integrate CBIR into automated planning workflow in future works, potentially through methods like the MetaPlanner framework.

Keywords: Automated treatment planning, Inverse Planning, Deep Learning, Knowledge based planning, Content based image retrieval

1. Introduction

The workflow for radiotherapy treatment planning typically involves an iterative, trial-and-error process for manually navigating trade-offs (Sethi 2018, Xing et al 1999). Treatment planning optimization contains multiple objectives, which are often conflicting. For this reason, no single plan can optimize performance on all objectives at once, and treatment planning can instead be conceptualized as navigating the set of Pareto optimal, nondominated solutions (Craft et al 2006, 2012, Huang et al 2021). In an effort to reduce the amount of active planning time in treatment planning, there has been growing interest in automated methods. Many of these methods (such as the MetaPlanner (MP) framework, the Expedited Constrained Hierarchical Optimization (ECHO) system, iCycle, etc.) can be interpreted as navigating the Pareto front while guided by some utility function (Huang et al 2022, Zarepisheh et al 2019, Breedveld et al 2012, Hussein et al 2018). While these
methods can obtain excellent results, even on complex clinical cases, a few potential limitations remain regarding the design of their utility functions.

Many utility functions used in treatment planning can be broadly categorized as either hand-crafted or data-driven. Hand-crafted utility functions are typically adapted from standard clinical protocols, as well as institutional protocols. In contrast, data-driven utility functions are fundamental to knowledge-based planning (KBP), and they typically involve training an end-to-end deep learning model to predict dose distributions or dose volume histograms (DVHs) (Hussein et al. 2018, Ma et al. 2019, Babier et al. 2021, Momin et al. 2021, Shen et al. 2020). In principle, using data-driven utility functions has the potential to more closely mimic the decision-making process of human planners. However, practically, there remain many limitations with current end-to-end KBP methods that need to be addressed.

Here, we highlight a few of the main limitations associated with current end-to-end KBP methods. First, end-to-end KBP models are often trained on limited datasets, and their performance may not generalize well between various institutions and protocols. As curating treatment planning data can be prohibitively expensive and public data remains relatively scarce (Babier et al. 2021), it may not be wise to rely on an end-to-end model entirely. Moreover, dose predictions from models following one protocol typically cannot be used for institutions that use different protocols. Consider, as an example, the Open-KBP dataset, which contains head and neck cases with planning target volumes (PTVs) prescribed doses of 70, 63, and 56 Gy. End-to-end models trained on this dataset cannot be easily applied for the purpose of planning cases using alternative protocols (e.g. using prescription doses of 70, 56, and 52 Gy) (National Cancer Institute (NCI) 2022).

Second, end-to-end KBP methods have no guarantees on the deliverability of their predicted dose distributions. These methods typically perform a pixel-wise regression of the dose distribution, and there are no inherent constraints on the predicted values. For this reason, the predicted dose distributions can be any or all of the following: infeasible, inefficient (not Pareto optimal), or non-compliant with clinical protocols.

In an attempt to address these limitations, this work proposes a content based image retrieval (CBIR) method that retrieves relevant treatment plans of patients from a database given a new patient’s anatomical information (i.e. medical images, contours, etc.). Unlike end-to-end methods, CBIR only utilizes deep learning for image representations. Image retrieval can be performed after filtering the database by specific institutions or protocols, which significantly improves generalizability. Similarly, instead of directly predicting the dose distribution, CBIR retrieves past patient plans, which satisfy the various criteria of being deliverable and protocol compliant.

2. Methods

2.1 Content Based Image Retrieval

Content-based image retrieval (CBIR) aims to search a database for images of similar content (i.e. anatomical information) to a query image. Figure 1 provides the overall CBIR workflow as applied to treatment planning. A database of previous treatment plans is first created and stored. This database contains each patient’s anatomical information, which includes their computed tomography (CT) images and relevant contours, as well as their dose distribution. After training the image encoding model, the CBIR method is supplied a new patient’s anatomical information, which it encodes into a latent space embedding that is compared to embeddings of other patients in the database. Image embeddings with the closest Euclidean distance are then retrieved from the database, and the corresponding dose distribution can be used in subsequent automated planning.

During deployment or real-world usage, the database is first filtered to contain plans with the relevant institution and clinical protocols. During all evaluations (i.e. benchmarking for this paper), the unfiltered database is used.

2.2 Image Encoding Model

Figure 1. Visualizes the workflow for CBIR. Given a new patient during treatment planning (i.e query image) the method searches a filtered database to retrieve similar images. The corresponding dose distribution can then be used in subsequent automated planning.
The main task of the image encoding model is to extract features from the provided images. Given images $X \in \mathbb{R}^D$, the goal is to learn an encoding function $f: \mathbb{R}^D \rightarrow \mathbb{R}^M$ that produces a continuous latent space embedding $z \in \mathbb{R}^M$. In this work we evaluate the image retrieval performance of five main categories of methods. Readers looking for model design inspiration may find previous reviews of alternative image retrieval tasks to be useful (Zin et al 2018, Dubey 2021, Latif et al 2019).

Prior to training the image encoding model, standard data pre-processing is applied to each patient’s images. First, each patient’s CT volume, segmentation mask, and corresponding dose distribution are resampled to the dimensions $d = 128 \times 128 \times 128$. The segmentation masks follow a label encoding scheme, containing the various planning target volumes (PTVs) and relevant organs-at-risk (OARs). After resampling, each CT volume is then clipped to a soft-tissue window and normalized.

This current work evaluates five main categories of image encoding models used for CBIR: (1) a vanilla autoencoder, (2) a variational autoencoder (VAE) (Zhao et al 2018), (3) a Siamese network with the triplet margin loss (Schroff et al 2015), (4) SimSiam (Chen and He 2020), and (5) a multitask Siamese network. For the encoder portion of all evaluated models, we utilize the same backbone convolutional neural network (CNN) architecture. Similarly, the latent space embedding size is empirically set to 1024 for all models.

2.2.1 Vanilla Autoencoder

![Figure 2a.](image1) Provides a schematic of the vanilla autoencoder model architecture.

The vanilla autoencoder consists of a standard CNN encoder and transposed convolution decoder. Figure 2a provides a schematic of the model.

2.2.2 Information Maximizing Variational Autoencoder

![Figure 2b.](image2) Provides a schematic of the InfoVAE model architecture.

The information maximizing variational autoencoder (InfoVAE) model is a generative model which uses an additional maximum mean discrepancy (Gretton et al 2006) objective, as proposed by Zhao et al. (Zhao et al 2018). As the InfoVAE is a generative model, the “μ” layer output is taken as the latent space embedding to avoid resampling during evaluation and deployment. Figure 2b provides a schematic of the InfoVAE architecture, and the loss function for the InfoVAE model is listed in Equation 1.

$$L_{\text{Info-VAE}} = -E_{x \sim p(x)}[\log p_\theta(x|z)] + (1 - \alpha)D_{KL}(q_\theta(z|x)||p(z)) + (\alpha + \lambda - 1)D_{MMD}(q_\theta(z|x)||p(z))$$

(1)

Here, the first term refers to the reconstruction loss, the second term refers to the KL divergence, and the third term refers to the maximum mean discrepancy. $p(z)$ refers to the prior, $q_\theta(z|x)$ refers to the variational posterior, $p_\theta(x|z)$ refers to the true posterior, and $\theta$ refers to the parameters of the network.

2.2.3 Siamese Network with a Triplet Loss Function

![Figure 2c.](image3) Provides a schematic of the Siamese network with a triplet loss function.

The Siamese network with a triplet loss (SNTL) is a classic method used for CBIR (Chechik et al 2010, Schroff et al 2015). The Siamese network refers to a network which contains duplicate encoders, where each shares parameters with its duplicates. The triplet loss function is provided in Equation 2. To construct triplets, we take a sample image from the dataset (i.e. anchor image). The positive image can then be sampled by taking another image from the same class as the anchor image, while the negative image refers to an image of...
Figure 2c provides a schematic of the SNTL model architecture.

\[
L_{\text{triplet}} = \max(d(X_a, X_p) - d(X_a, X_n) + \text{margin}, 0) \\
d(X_1, X_2) = ||X_1 - X_2||_2
\]  

(2)

2.2.4 Simple Siamese Network

The simple Siamese (SimSiam) network (Chen and He 2020) is a recent representation learning method that extends on previous state of the art methods like SimCLR (Chen et al 2020a) and BYOL (Grill et al 2020). SimSiam uses a stop-gradient to learn meaningful representations without the use of negative sample pairs, large batches, or momentum encoders. As these are usually difficult to obtain, utilizing an approach like SimSiam can be more practical than other recent representation learning methods. Figure 2d provides a schematic of the SimSiam model, and the loss function is listed in Equation 3. In this work, we use the corresponding dose distribution of each patient as the transformed image instead of a geometrically transformed image.

Here, \( p_1 = h(f(x_1)) \) refers to the output of the predictor \( h \), and \( z_2 \) refers to the embedding of the transformed image.

\[
L_{\text{SimSiam}} = -\frac{p_1}{||p_1||_2} \cdot \frac{z_2}{||z_2||_2}
\]  

(3)

2.2.5 Multitask Siamese Network

The multitask siamese network (MSN) combines many of the previously mentioned approaches. Due to the small dataset size of this study, the MSN hopes to improve generalization by utilizing information from the reconstruction task, SimSiam embedding task, and the triplet loss ranking task. Figure 2e provides a schematic of the MSN model, and the loss function is also provided in Equation 4.

\[
L_{\text{multitask}} = L_{\text{recon}} + \beta L_{\text{SimSiam}} + \gamma L_{\text{Triplet}}
\]  

(4)

2.3 Dataset

The dataset used in this current study contains 405 cases composed of public data (i.e. OpenKBP) and private data, collected as part of clinical workflow. The body sites included in this dataset are prostate and head and neck, with either volumetric modulated arc therapy (VMAT) or intensity modulated radiation therapy (IMRT) used for treatment.

For evaluation, all cases in the dataset were manually classified according to the following four criteria for a total of 32 classes:
1. Is the case prostate or head and neck?
2. Is there one PTV dose level or multiple PTV dose levels?
3. Is the primary PTV small or large?
4. How is the primary PTV located (i.e. left, right, center, or bilateral)?

Figure 3 provides a visualization of the workflow taken to classify patients in the dataset. Cases were additionally split into 235 in the training phase, 43 in the validation phase, and 127 in the testing phase. All source code for this project has been made available on github (https://github.com/chh105/MetaPlanner/tree/main/cbir).

3. Results

3.1 Evaluation

Performance of the included image retrieval methods was evaluated for three aspects: retrieval performance, clustering performance, and qualitative performance. We begin by retrieving $k$ images from the database that have embeddings closest to that of the query image. Retrieval performance can then be evaluated using standard metrics like the classification accuracy, precision, recall, and F-score at rank $k$ (Mogotsi 2010). The definitions for these evaluation metrics are listed in Table 1. Moreover, clustering performance is then evaluated using standard metrics like the cluster homogeneity, completeness, v-measure, adjusted Rand index, and adjusted mutual information (Rosenberg and Hirschberg 2007, Hubert and Arabie 1985, Steinley 2004, Strehl and Ghosh 2003). Lastly, qualitative performance is evaluated by visually inspecting the retrieval results for example query patients.

![Image of retrieval metrics](Figure 4. Plot of the retrieval metrics for the top-k images. For all retrieval metrics, best performance was achieved using the multitask Siamese network.)

Table 1. Definitions for various retrieval (multiclass) evaluation metrics.

| Metric            | Definition                                                                 |
|-------------------|-----------------------------------------------------------------------------|
| **Accuracy@k**    | $f(k) = \frac{1}{{N_{classes}}} \sum_{i=1}^{N_{classes}} \frac{tp + tn}{tp + fp + fn + tn}$ |
| **Precision@k**   | $f(k) = \frac{1}{{N_{classes}}} \sum_{i=1}^{N_{classes}} \frac{tp}{tp + fp}$                     |
| **Recall@k**      | $f(k) = \frac{1}{{N_{classes}}} \sum_{i=1}^{N_{classes}} \frac{tp}{tp + fn}$                     |
| **F1-score@k**    | $f(k) = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$  |

Retrieval Score

$s = \sum_{k=1}^{N} f(k) \cdot \left(\frac{1}{2}\right)^k$

- $tp$ := true positives, $tn$ := true negatives
- $fp$ := false positives, $fn$ := false negatives

3.2 Image Retrieval Performance

We first evaluate the image retrieval performance of the candidate image encoding models using the metrics in Table 1. Figure 4 plots accuracy, precision, recall, and F-score as functions of $k$. Ground-truth labels for each patient are provided by following the procedure described in Section 2.3. For each metric, we can compute a simple retrieval scoring function by applying an exponential weighting to each retrieval metric. Here, greater emphasis is placed on small values of $k$, as only the most relevant retrieved plans would be
used to inform subsequent treatment planning. Values of the retrieval scoring functions are listed in Table 2.

For all retrieval scores, MSN achieves the best performance. The second-best performance for all retrieval scores is achieved by the SNTL method, followed by the more recent SimSiam method. Performances for the vanilla autoencoder and Info-VAE were comparable, suggesting that the variational loss component may not be entirely useful for our image retrieval task.

3.3 Clustering Performance

Clustering performance of the candidate methods is additionally evaluated using the cluster homogeneity, completeness, V-measure, adjusted Rand index, and adjusted mutual information. Each score is computed on the latent space embeddings produced by the candidate methods, where ground truth labels are provided following the procedure detailed in Section 2.3 and prediction labels are computed for \( k = 1 \).

Cluster homogeneity assesses the ability to create clusters that contain only members of a single class. Cluster homogeneity is bounded between \([0,1]\), and performance of an ideal method approaches a cluster homogeneity of 1. Of the benchmarked encoders, the top three performers for cluster homogeneity were the MSN, SNTL, and Info VAE models.

Cluster completeness assesses the ability to assign all members of a class to the same cluster. Cluster completeness is bounded between \([0,1]\), and performance of an ideal method approaches a cluster completeness of 1. Of the benchmarked encoders, the top three performers for cluster completeness were the MSN, SNTL, and Info VAE models.

V-measure is computed as the harmonic mean of cluster homogeneity and completeness. V-measure is bounded between \([0,1]\), and performance of an ideal method also approaches a value of 1. Of the benchmarked encoders, the top

Table 2. Performance of the candidate encoding models is evaluated in regards to retrieval and clustering. Retrieval scores are computed using an exponential weighting of each metric as a function of \( k \). Cluster metrics are computed using standard formulas without an exponential weighting.

| Model                                      | Accuracy Retrieval Score | Precision Retrieval Score | Recall Retrieval Score | F\(_1\) Retrieval Score | Homogeneity | Completeness | V-measure | Adjusted Rand Index | Adjusted Mutual Info. |
|--------------------------------------------|---------------------------|----------------------------|------------------------|--------------------------|-------------|--------------|-----------|---------------------|-----------------------|
| Multitask Siamese Network                  | 1.23                      | 1.04                       | 1.03                   | 1.02                     | 0.683       | 0.679        | 0.681     | 0.516               | 0.593                 |
| Info VAE                                   | 0.70                      | 0.58                       | 0.57                   | 0.52                     | 0.438       | 0.427        | 0.432     | 0.275               | 0.278                 |
| Siamese Network (Triplet Loss)             | 1.11                      | 0.96                       | 0.97                   | 0.95                     | 0.671       | 0.653        | 0.662     | 0.437               | 0.571                 |
| SimSiam (Cosine Similarity)                | 0.78                      | 0.80                       | 0.59                   | 0.60                     | 0.412       | 0.422        | 0.417     | 0.247               | 0.265                 |
| Vanilla Autoencoder                        | 0.72                      | 0.62                       | 0.56                   | 0.56                     | 0.372       | 0.375        | 0.374     | 0.194               | 0.204                 |

Figure 5. Provides a visualization of the latent space embeddings of the query images computed by each of the candidate encoding models. Only body site classification labels are included here for visualization purposes.
three performers for V-measure were the MSN, SNTL, and Info VAE models.

The adjusted Rand index measures the similarity between ground truth class assignments and those of the clustering method, adjusted for chance groupings. The adjusted Rand index is bounded between \([-1, 1]\), and performance of an ideal method approaches a value of 1. Of the benchmarked encoders, the top three performers for the adjusted Rand index were the MSN, SNTL, and Info VAE models.

Finally, the adjusted mutual information measures the agreement between ground truth class assignments and those of the clustering method, adjusted for chance groupings. The adjusted mutual information is bounded between \([0, 1]\), and performance of an ideal method approaches a value of 1. Of the benchmarked encoders, the top three performers for the adjusted mutual information were the MSN, SNTL, and Info VAE models.

Figure 5 shows the latent space embeddings for the query set, with the body site categorizations provided for visualization purposes (ground truth labels for all evaluations are computed following Section 2.3). Here, embeddings for the MSN are substantially more distinct and grouped than those of the other candidate models.

3.4 Qualitative Performance

A qualitative comparison of retrieved images for an example query image is provided in Figure S1 of the Supplemental Materials. This example query image is a head and neck case, has multiple PTV levels, and has a large primary PTV located on the right-side of the patient. Both the MSN and SNTL models retrieve patients of the same classification from the database. Moreover, the retrieved patient for the MSN model is more anatomically similar to the query than that of the SNTL model. The remaining models did not retrieve patients from the database of the same classification as the query image. For all evaluations in this current work, retrieval is performed on the unfiltered database. However, during practical deployment, the database will first be filtered by the relevant classification.

4. Discussion

Here, a CBIR framework is used to retrieve treatment plans from a database and inform subsequent automated planning. The proposed CBIR framework compares the latent space embeddings of a query image to those of images in a database for the purpose of image retrieval. To produce latent space embeddings, we evaluate various encoding models in regards to retrieval performance, clustering performance, and visual quality.

As discussed in previous sections, traditional end-to-end KBP methods are often trained using limited datasets and may not generalize well to different institutions and protocols. Instead of relying completely on learning-based approaches, CBIR only uses an encoding model to produce latent space embeddings of each patient’s anatomical information. Consequently, errors accrued from poor model performance have less of an impact on the predicted dose distributions than those accrued when performing an end-to-end approach.

During CBIR deployment, we can additionally filter the database to contain only relevant plans from specified institutions or protocols, thereby ensuring that the retrieved treatment plans are protocol compliant and deliverable.

The proposed CBIR framework retrieves relevant treatment plans from a database and can be utilized in any pipeline that would otherwise incorporate end-to-end KBP. Specifically, the retrieved dose distributions can be used in methods which directly optimize machine parameters through dose mimicking (Eriksson and Zhang n.d., McIntosh et al. 2017, Mahmood et al. 2018). They can be alternatively used in modular methods like the MetaPlanner framework (Huang et al. 2022), which optimize treatment planning hyperparameters and can be more robust than direct dose mimicking.

In this work, various candidate encoder models were evaluated to determine viable options for treatment planning CBIR. Of the evaluated methods, the multitask Siamese network consistently performed the best in regards to retrieval performance, clustering performance, and visual quality. The dataset used in this study includes a total of 405 cases. Though this may be considered sizeable in the context of medical data, it certainly cannot compare to datasets used routinely in computer vision (Deng et al. 2009, Lecun et al. 1998). Given the relatively small dataset size used in this current study, the multitask model manages to outperform its alternatives by incorporating additional loss function terms to reduce overfitting. This is evident when observing the performance of methods like SNTL, SimSiam, or the vanilla autoencoder, which individually do not perform as well as the multitask model.

This current study is subject to some limitations. First, while several candidate methods for encoding images were evaluated here, there may certainly exist better performing encoding models that were not tested. Second, due to data availability, we were not able to evaluate other body sites such as lung data, liver data, etc.

External beam radiation therapy is a highly popular treatment modality (Bilimoria et al. 2008). Recently, there has been growing interest in developing automated methods for the radiotherapy pipeline. Deep learning has generally been successful in performing radiotherapy tasks like segmentation, outcome prediction, etc. (Boldrini et al. 2019, Liang et al. 2021, Yuan et al. 2019, Chen et al. 2020b, Nomura et al. 2020, Dong and Xing 2020, Pastor-Serrano and Perkó 2021, 2022), and applying learning based methods to treatment planning also has potential. In future works, we hope to apply the proposed CBIR method directly to automated planning, potentially through frameworks like MetaPlanner (Huang et al. 2022).
Similarly, we plan to address some of the mentioned limitations of this current study by evaluating additional CBIR encoding models and utilizing data from additional body sites.

5. Conclusion

In this work, we introduced a CBIR method to inform subsequent treatment planning. The proposed workflow addresses some key limitations present in traditional end-to-end KBP methods, including generalizability, deliverability, and protocol compliance of predicted dose distributions. To determine a viable encoding model for CBIR, we evaluated several methods ranging from the Info-VAE to Siamese networks with various loss functions to a multitask network that combines tasks from other candidate approaches. Our results indicate that the multitask encoding model consistently provides the best performance when evaluated in regards to retrieval performance, clustering performance, and visual quality.

Acknowledgements

This research was supported by the National Institutes of Health (NIH) under Grants 1R01 CA176553, R01CA227713, and T32EB009653, Varian Medical Systems, as well as a faculty research award from Google Inc.

References

Babier A, Zhang B, Mahmood R, Moore K L, Purdie T G, McNiven A L and Chan T C Y 2021 OpenKBP: The open-access knowledge-based planning grand challenge Med. Phys. 48 5549–61

Bilimoria K Y, Stewart A K, Winchester D P and Ko C Y 2008 The National Cancer Data Base: A Powerful Initiative to Improve Cancer Care in the United States Ann. Surg. Oncol. 15 683–90

Boldrini L, Bibault J-E, Masciocchi C, Shen Y and Bittner M-I 2019 Deep Learning: A Review for the Radiation Oncologist Front. Oncol. 9 Online: https://www.frontiersin.org/article/10.3389/fonc.2019.00977

Breedveld S, Storchi P R M, Voet P W J and Heijmen B J M 2012 iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans Med. Phys. 39 951–63

Chechik G, Sharma V, Shalit U and Bengio S 2010 Large Scale Online Learning of Image Similarity Through Ranking J. Mach. Learn. Res. 11 1109–35

Chen T, Kornblith S, Norouzi M and Hinton G 2020a A Simple Framework for Contrastive Learning of Visual Representations ArXiv200205709 Cs Stat Online: http://arxiv.org/abs/2002.05709

Chen X and He K 2020 Exploring Simple Siamese Representation Learning ArXiv20110566 Cs Online: http://arxiv.org/abs/2011.10566

Chen Y, Xing L, Yu L, Bagshaw H P, Buyounouski M K and Han B 2020b Automatic intraprostatic lesion segmentation in multiparametric magnetic resonance images with proposed multiple branch UNet Med. Phys. 47 6421–9

Craft D L, Halabi T F, Shih H A and Bortfeld T R 2006 Approximating convex Pareto surfaces in multijobjective radiotherapy planning Med. Phys. 33 3399–407

Craft D L, Hong T S, Shih H A and Bortfeld T R 2012 Improved Planning Time and Plan Quality Through Multicriteria Optimization for Intensity-modulated Radiotherapy Int. J. Radiat. Oncol. Biol. Phys. 82 e83–90

Deng J, Dong W, Socher R, Li L-J, Li K and Fei-Fei L 2009 ImageNet: A large-scale hierarchical image database 2009 IEEE Conference on Computer Vision and Pattern Recognition 2009 IEEE Conference on Computer Vision and Pattern Recognition pp 248–55

Dong P and Xing L 2020 Deep DoseNet: a deep neural network for accurate dosimetric transformation between different spatial resolutions and/or different dose calculation algorithms for precision radiation therapy Phys. Med. Amptextbackslashmathsemicolon Biol. 65 035010

Dubey S R 2021 A Decade Survey of Content Based Image Retrieval using Deep Learning IEEE Trans. Circuits Syst. Video Technol. 1–1

Eriksson O and Zhang T Robust automated radiation therapy treatment planning using scenario-specific dose prediction and robust dose mimicking Med. Phys. n/a Online: https://onlinelibrary.wiley.com/doi/abs/10.1002/mp.15622

Gretton A, Borgwardt K, Rasch M, Schölkopf B and Smola A 2006 A Kernel Method for the Two-Sample-Problem Advances in Neural Information Processing Systems vol 19 (MIT Press) Online: https://papers.nips.cc/paper/2006/hash/e9f22eda3d9c55a0d89c98d6c54b53e-Abstract.html

Grill J-B, Strub F, Alcâê F, Tallec C, Richemond P H, Buchatskaya E, Doersch C, Pires B A, Guo Z D, Azar M G, Piot B, Kavukcuoglu K, Munos R and Valko M 2020 Bootstrap your own latent: A new approach to self-supervised Learning ArXiv200607733 Cs Stat Online: http://arxiv.org/abs/2006.07733

Huang C, Nomura Y, Yang Y and Xing L 2022 Meta-optimization for fully automated radiation therapy treatment planning Phys. Med. Biol. Online: http://iopscience.iop.org/article/10.1088/1361-6560/ac5672
Huang C, Yang Y, Panjwani N, Boyd S and Xing L 2021 Pareto Optimal Projection Search (POPS): Automated Radiation Therapy Treatment Planning by Direct Search of the Pareto Surface IEEE Trans. Biomed. Eng. PP

Hubert L and Arabie P 1985 Comparing partitions J. Classif. 2 193–218

Hussein M, Heijmen B J M, Verellen D and Nisbet A 2018 Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations Br. J. Radiol. 91 20180270

Latif A, Rasheed A, Sajid U, Ahmed J, Ali N, Ratyal N I, Zafar B, Dar S H, Sajid M and Khalil T 2019 Content-Based Image Retrieval and Feature Extraction: A Comprehensive Review Math. Probl. Eng. 2019 e9658350

Lecun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition Proc. IEEE 86 2278–324

Liang X, Bibault J-E, Leroy T, Escande A, Zhao W, Chen Y, Buyyounouski M K, Hancock S L, Bagshaw H and Xing L 2021 Automated contour propagation of the prostate from pCT to CBCT images via deep unsupervised learning Med. Phys. 48 1764–70

Ma M, Kovalchuk N, Buyyounouski M K, Xing L and Yang Y 2019 Incorporating dosimetric features into the prediction of 3D VMAT dose distributions using deep convolutional neural network Phys. Med. Biol. 64 125017

Mahmood R, Babier A, McNiven A, Diamant A and Chan T C Y 2018 Automated Treatment Planning in Radiation Therapy using Generative Adversarial Networks ArXiv180706489 Phys. Stat. Online: http://arxiv.org/abs/1807.06489

McIntosh C, Welch M, McNiven A, Jaffray D A and Purdie T G 2017 Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method Phys. Med. Biol. 62 5926–44

Mogetsi I C 2010 Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze: Introduction to information retrieval Inf. Retr. 13 192–5

Momin S, Fu Y, Lei Y, Roper J, Bradley J D, Curran W J, Liu T and Yang X 2021 Knowledge-based radiation treatment planning: A data-driven method survey J. Appl. Clin. Med. Phys. 22 16–44

National Cancer Institute (NCI) 2022 A Randomized Phase II/III Trial of De-Intensified Radiation Therapy for Patients With Early-Stage, P16-Positive, Non-Smoking Associated Oropharyngeal Cancer (clinicaltrials.gov) Online: https://clinicaltrials.gov/ct2/show/NCT03952585

Nomura Y, Wang J, Shirato H, Shimizu S and Xing L 2020 Fast spot-scanning proton dose calculation method with uncertainty quantification using a three-dimensional convolutional neural network Phys. Med. Amptextbackslashmathsemicolon Biol. 65 215007

Pastor-Serrano O and Perkó Z 2021 Learning the Physics of Particle Transport via Transformers (arXiv) Online: http://arxiv.org/abs/2109.03951

Pastor-Serrano O and Perkó Z 2022 Millisecond speed deep learning based proton dose calculation with Monte Carlo accuracy Phys. Med. Amptextbackslashmathsemicolon Biol. 67 105006

Rosenberg A and Hirschberg J 2007 V-Measure: A Conditional-Entropy-Based External Cluster Evaluation Measure Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) CoNLL-EMNLP 2007 (Prague, Czech Republic: Association for Computational Linguistics) pp 410–20 Online: https://aclanthology.org/D07-1043

Schroff F, Kalenichenko D and Philbin J 2015 FaceNet: A Unified Embedding for Face Recognition and Clustering 2015 IEEE Conf. Comput. Vis. Pattern Recognit. CVPR 815–23

Sethi A 2018 Khan’s Treatment Planning in Radiation Oncology, 4th Edition. Editor: Faiz M. Khan, John P. Gibbons, Paul W. Sperduto. Lippincott Williams & Wilkins (Wolters Kluwer), Philadelphia, PA, 2016. 648 pp. Price: $215.99. ISBN: 9781469889979. (Hardcover) (*) Med. Phys. 45 2351–2351

Shen C, Nguyen D, Chen L, Gonzalez Y, McBeth R, Qin N, Jiang S B and Jia X 2020 Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning Med. Phys.

Steinley D 2004 Properties of the Hubert-Arable Adjusted Rand Index Psychol. Methods 9 386–96

Strehl A and Ghosh J 2003 Cluster ensembles --- a knowledge reuse framework for combining multiple partitions J. Mach. Learn. Res. 3 583–617

Xing L, Li J G, Donaldson S, Le Q T and Boyer A L 1999 Optimization of importance factors in inverse planning Phys. Med. Biol. 44 2525–36
Yuan Y, Qin W, Guo X, Buyyounouski M, Hancock S, Han B and Xing L 2019 Prostate Segmentation with Encoder-Decoder Densely Connected Convolutional Network (Ed-Densenet) 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) pp 434–7

Zarepisheh M, Hong L, Zhou Y, Oh J H, Mechalakos J G, Hunt M A, Mageras G S and Deasy J O 2019 Automated intensity modulated treatment planning: The expedited constrained hierarchical optimization (ECHO) system Med. Phys. 46 2944–54

Zhao S, Song J and Ermon S 2018 InfoVAE: Information Maximizing Variational Autoencoders ArXiv170602262 Cs Stat Online: http://arxiv.org/abs/1706.02262

Zin N A M, Yusof R, Lashari S A, Mustapha A, Senan N and Ibrahim R 2018 Content-Based Image Retrieval in Medical Domain: A Review J. Phys. Conf. Ser. 1019 012044
Figure S1. Provides a visualization of the retrieval results for an example query image. The multitask Siamese network retrieves a patient of the same class as the query image that is also anatomically similar. The SNTL model also retrieves a patient of the same class, though the retrieved image has less anatomical similarity to the query. For this query example, the remaining methods do not retrieve an image of the same class.