Learning image representations for content-based image retrieval of radiotherapy treatment plans

Huang, Charles; Vasudevan, Varun; Pastor-Serrano, Oscar; Islam, Md Tauhidul; Nomura, Yusuke; Dubrowski, Piotr; Wang, Jen Yeu; Schulz, Joseph B.; Yang, Yong; More Authors

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

PAPER

Learning image representations for content-based image retrieval of radiotherapy treatment plans

View the article online for updates and enhancements.

Physics in Medicine & Biology

PAPER

Learning image representations for content-based image retrieval of radiotherapy treatment plans

Charles Huang[1,][*] ⓘ, Varun Vasudevan[2] ⓘ, Oscar Pastor-Serrano[3,4], Md Tauhidul Islam[3] ⓘ,
Yusuke Nomura[3] ⓘ, Piotr Dubrowski[3], Jen-Yeu Wang[3], Joseph B Schulz[3], Yong Yang[3] and Lei Xing[3]

1 Department of Bioengineering, Stanford University, Stanford, United States of America
2 Institute for Computational and Mathematical Engineering, Stanford University, Stanford, United States of America
3 Department of Radiation Oncology, Stanford University, Stanford, United States of America
4 Department of Radiation Science and Technology, Delft University of Technology, The Netherlands
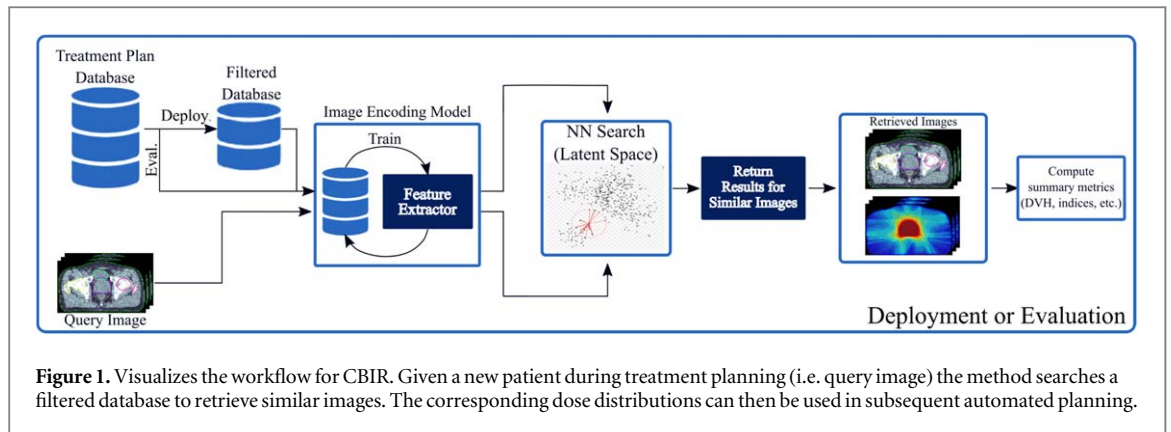* Author to whom any correspondence should be addressed.

E-mail: chh105@stanford.edu

Abstract

Objective. In this work, we propose a content-based image retrieval (CBIR) method for retrieving dose distributions of previously planned patients based on anatomical similarity. Retrieved dose distributions from this method can be incorporated into automated treatment planning workflows in order to streamline the iterative planning process. As CBIR has not yet been applied to treatment planning, our work seeks to understand which current machine learning models are most viable in this context. *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. 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 image sets and clinical image sets 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.* 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.

1. Introduction

1.1. Background

The workflow for radiotherapy treatment planning typically involves an iterative, trial-and-error process for manually navigating trade-offs (Xing *et al* 1999, Sethi 2018). 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 active planning times 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 (Breedveld *et al* 2012, Hussein *et al* 2018, Zarepisheh *et al* 2019, Huang *et al* 2022). Such utility functions are typically designed around clinical protocols and incorporate various dose metrics to gauge treatment plan quality.

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 distributions can then be used in subsequent automated planning.

At the same time, with recent advancements in machine learning research, interest in data-driven automated treatment planning approaches has begun to surge as well. Many of these data-driven approaches have elected to replace conventional utility functions entirely, instead using an end-to-end deep learning model that produces predictions of dose distributions or dose volume histograms (DVHs) (Hussein *et al* 2018, Ma *et al* 2019, Shen *et al* 2020, Babier *et al* 2021, Momin *et al* 2021). In principle, data-driven methods attempt to capture the collective expertise of numerous treatment planners in their model predictions. Yet, due to the relative data scarcity in medical imaging and treatment planning datasets, data-driven methods can have many drawbacks in practice. Thus, it may not be prudent to rely entirely on end-to-end machine learning models. As a compromise, we propose a cascaded approach that first performs content-based image retrieval (CBIR) and then subsequently automated treatment planning.

1.2. Content-based image retrieval

Content-based image retrieval refers to a category of methods that retrieve relevant images from a database based on analysed content of a query image. In the context of treatment planning, the problem can be framed as searching a database for relevant treatment plan information given a new patient's anatomical information (i.e. medical images, contours, etc), called the query image. During deployment of a CBIR method in a clinical setting, the new patient's treatment plan has obviously not been created yet, so CBIR involves analysis of anatomical information to find the most relevant previously treated patient(s). The dose information of one or more previously treated patients that have been deemed relevant can then be used in subsequent automated planning.

CBIR uses a machine learning model to create latent space representations of the query image and images from the database (Zin *et al* 2018, Latif *et al* 2019, Dubey 2021). After computing a distance function (i.e. Euclidean distance) between the query image representation and representations of the database images, we can then sort by the closest distances (Nearest neighbour search) and return the k closest images to the query (Top-k images).

Unlike end-to-end methods that use machine learning predictions for the entire workflow, CBIR only utilizes deep learning for image representations and has several potential advantages. For instance, CBIR methods can be more easily adapted to clinical protocols beyond the ones seen during training. Adapting CBIR to new protocols or guidelines simply involves filtering the database that the retrieval method selects from to only include patients that follow those new desired protocols. Figure 1 provides a visualization of database filtering when deploying CBIR in a clinical setting. For the evaluation or benchmarking purposes of this manuscript, all results will be provided for an unfiltered database.

The end-to-end machine learning methods used for treatment planning have the potential to reap the benefits of amortized inference for improved efficiency as compared to conventional methods. However, due to the practical limitations of acquiring large, heterogeneous datasets in treatment planning, using end-to-end methods may not be advisable. CBIR provides a compromise between end-to-end machine learning methods and conventional automated planning methods. To the best of our knowledge, CBIR has not previously been applied to treatment planning. As such, this work seeks to answer fundamental questions around selecting viable machine learning models for CBIR. Here, we compare several potential image encoding models for CBIR and describe their methodologies below.

2. Methods

2.1. Content-based image retrieval

Content-based image retrieval 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, the query image, which it encodes into a latent space embedding that is compared to embeddings of other patients in the database. Image embeddings (i.e. one-dimensional vector representations of each image in the latent space) with the closest Euclidean distance are then retrieved from the database (Nearest neighbour search), 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 in this paper, the unfiltered database is used.

2.2. Image encoding models

The main task of the image encoding model is to extract features from the provided images. Given images $X \in R^D$, the goal is to learn an encoding function $f: R^D \rightarrow R^M$ that produces a continuous latent space embedding $z \in R^M$. Here, X refers to a multichannel volume which consists of the CT and contours for each patient, and X_i refers to an input from the branch i, numbered in ascending order for branches going from top to bottom (i.e. $i = 2$ refers to the second branch from the top). For the methods that utilize contrastive learning (i.e. SimSiam and the multitask Siamese network), which are presented in later sections, the input during training also includes a channel for the dose distribution. During deployment and evaluation, the input to all models only includes the CT and contours. The embedding z refers to the one-dimensional vector representation of images in the latent space. 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, Latif *et al* 2019, Dubey 2021).

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 (with labels ranging from 1 to 4), 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 (400 HU width and 0 HU level) and normalized. To keep consistent with common convention in Siamese networks, we interchangeably refer to the input image as the anchor image.

This current work evaluates five main categories of image encoding models used for CBIR: (1) a vanilla autoencoder (Goodfellow *et al* 2016), (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 (Caruana 1997, Schroff *et al* 2015, Chen and He 2020). For the encoder portion of all evaluated models, we utilize the same backbone convolutional neural network (CNN) architecture (consisting of multiple convolution, GroupNorm, and LeakyReLU blocks). Similarly, the latent space embedding vector has a size that is empirically set to 1024 for all models.

2.2.1. Vanilla autoencoder

The vanilla autoencoder consists of a standard CNN encoder and a transposed convolution decoder (Goodfellow *et al* 2016). Figure 2(a) provides a schematic of the model.

2.2.2. Information maximizing variational autoencoder

The information maximizing variational autoencoder (Info-VAE) model is a generative model which uses an additional maximum mean discrepancy (Gretton *et al* 2006) objective, as proposed by Zhao *et al* (2018). As the Info-VAE is a generative model, the typical embedding vector z is a random variable sample using the reparameterization trick (Kingma and Welling 2013, Zhao *et al* 2018). In order to get a deterministic output, we follow the common practice of using the 'μ' layer output as the latent space embedding vector instead of the vector z.

Figure 2(b) provides a schematic of the InfoVAE architecture, and the loss function for the InfoVAE model is listed in equation (1)

Figure 2. (a) A schematic of the vanilla autoencoder model architecture. (b) A schematic of the information-maximizing variational autoencoder model architecture. Outputs of the encoder are vectors describing the mean and variance of the latent space distributions. (c) A schematic of the Siamese network with a triplet loss function. (d) A schematic of the SimSiam network, which uses a stop gradient and the cosine distance. Here the transformed image refers to the dose distribution of patients during model training. Only the anchor image is required for deployment and evaluation. (e) A schematic of the multitask Siamese network, which combines tasks from many of the previously discussed approaches.

$$
\begin{aligned}
L_{\mathrm{Info-VAE}} = {} & -E_{z \sim q(z|x)}[\log p_\theta(x|z)] \\
& + (1 - \alpha)D_{\mathrm{KL}}(q_\theta(z|x) \| p(z)) \\
& + (\alpha + \lambda - 1)D_{\mathrm{MMD}}(q_\theta(z|x) \| p(z)).
\end{aligned}
\tag{1}
$$

Here, the first term refers to the reconstruction loss (typically mean squared error), the second term D_{KL} refers to the Kullback–Leibler (KL) divergence, and the third term D_{MMD} refers to the maximum mean discrepancy (Gretton *et al* 2006). $p(z)$ refers to the prior, $q_\theta(z|x)$ refers to the variational posterior, $p_\theta(x|z)$ refers

to the true posterior, α and λ are hyperparameters controlling the amount of regularization, and θ refers to the parameters of the network (Kingma and Welling 2013, Zhao *et al* 2018).

2.2.3. 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 (see the Dataset section) as the anchor image, while the negative image refers to an image of a different class than the anchor image. Figure 2(c) provides a schematic of the SNTL model architecture. The triplet loss function computes a distance between the embeddings for the anchor image z_1 and positive image z_p, as well as a distance between the embeddings for the anchor image and negative image z_n. The goal is then to make the distance between the anchor and positive embeddings at least some margin smaller than the distance between the anchor and negative embeddings.

$$L_{\text{Triplet}} = \max\left(\|z_1 - z_p\|_2 - \|z_1 - z_n\|_2 + \text{margin}, 0\right) \tag{2}$$

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 2(d) provides a schematic of the SimSiam model, and the loss function is listed in equation (3). In order to incorporate information from the dose distributions during model training, we set the transformed image as a multichannel input that uses the dose distribution and contour information. This training scheme is also used by the multitask approach described below

$$L_{\text{SimSiam}} = -\frac{p_1}{\|p\|_{12}} \cdot \frac{z_2}{\|z_2\|_2}. \tag{3}$$

Here, $p_1 = h(f(x_1))$ refers to the output of the predictor (h), z_1 refers to the embedding of the anchor image, and z_2 refers to the embedding of the transformed image.

2.2.5. Multitask siamese network

Following the typical multitask learning scheme (Caruana 1997), the multitask Siamese network (MSN) combines many of the previously mentioned approaches. Due to the small dataset size of this study, the MSN attempts to improve generalization by utilizing information from the reconstruction task, SimSiam embedding task, and the triplet loss task. Figure 2(e) provides a schematic of the MSN model,

$$L_{\text{multitask}} = L_{\text{recon}} + \beta L_{\text{SimSiam}} + \gamma L_{\text{Triplet}} \tag{4}$$

and the loss function is provided in equation (4).

In this study, β and γ were empirically set to values of $1e-2$ and $1e-1$, respectively.

2.3. Dataset

The dataset used in this study contains 405 cases composed of public data (OpenKBP) (Babier *et al* 2021) and institutional data, which were 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. Which body site does the case belong to (i.e. prostate or head and neck)?

2. How many target levels are there?

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 split into 235 in the training phase, 43 in the validation phase, and 127 in the testing phase. Small or large labels were

Figure 3. Visualizes the workflow for classifying an example patient for benchmark evaluation purposes.
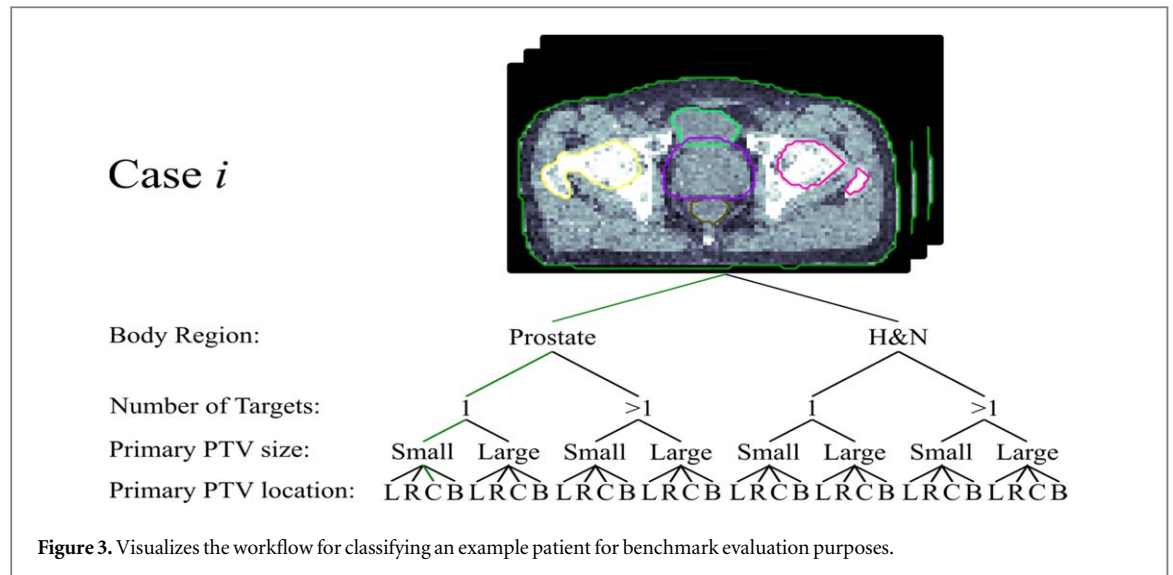
Table 1. Definitions for various retrieval (multiclass) evaluation metrics. Each metric is computed for the top-k images returned by the retrieval systems.
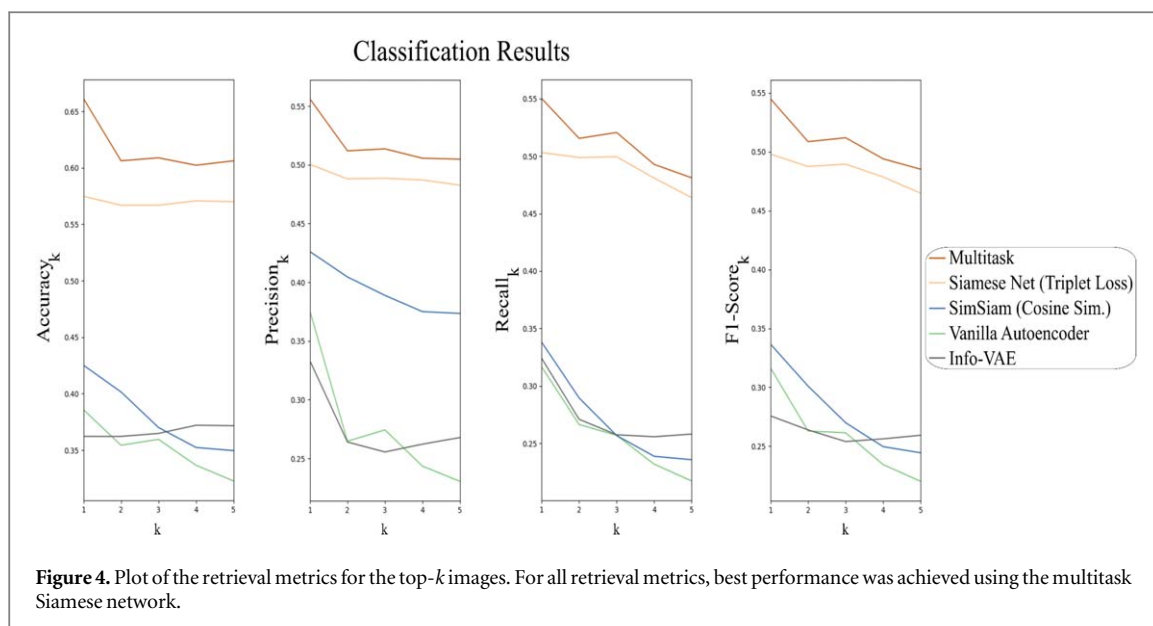
	Definition
Accuracy@k	$f(k) = \dfrac{1}{N_{\text{classes}}} \displaystyle\sum_{j=1}^{N_{\text{classes}}} \dfrac{tp + tn}{tp + fp + fn + tn}$
Precision@k	$f(k) = \dfrac{1}{N_{\text{classes}}} \displaystyle\sum_{i=1}^{N_{\text{classes}}} \dfrac{tp}{tp + fp}$
Recall@k	$f(k) = \dfrac{1}{N_{\text{classes}}} \displaystyle\sum_{i=1}^{N_{\text{classes}}} \dfrac{tp}{tp + fn}$
F_1 − score@k	$f(k) = \dfrac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$
Retrieval score	$s = \displaystyle\sum_{k=1}^{n} f(k) \cdot \left(\dfrac{1}{2}\right)^{k}$
	tp: = true positives, tn: = true negatives
	fp: = false positives, fn: = false negatives

assigned to cases if their primary PTV was smaller or larger than the median volume measurement for the dataset. All source code for this project has been made available on github https://github.com/chh105/MetaPlanner/tree/main/cbir.

2.4. 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_1-score of the top-k retrieved images (Mogotsi 2010). In this study, we show results for k ranging from 1 to 5, though other ranges of k may also be used. 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(Hubert and Arabie 1985, Strehl and Ghosh 2003, Steinley 2004, Rosenberg and Hirschberg 2007). Lastly, qualitative performance is evaluated by visually inspecting the retrieval results for example query patients.

Cluster homogeneity (Rosenberg and Hirschberg 2007) 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. Cluster completeness (Rosenberg and Hirschberg 2007) 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. V-measure (Rosenberg and Hirschberg 2007) 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. The adjusted Rand index (Hubert and Arabie 1985) measures the similarity between ground truth class assignments and those of the

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.

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. Finally, the adjusted mutual information (Strehl and Ghosh 2003) 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.

3. Results

3.1. 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.
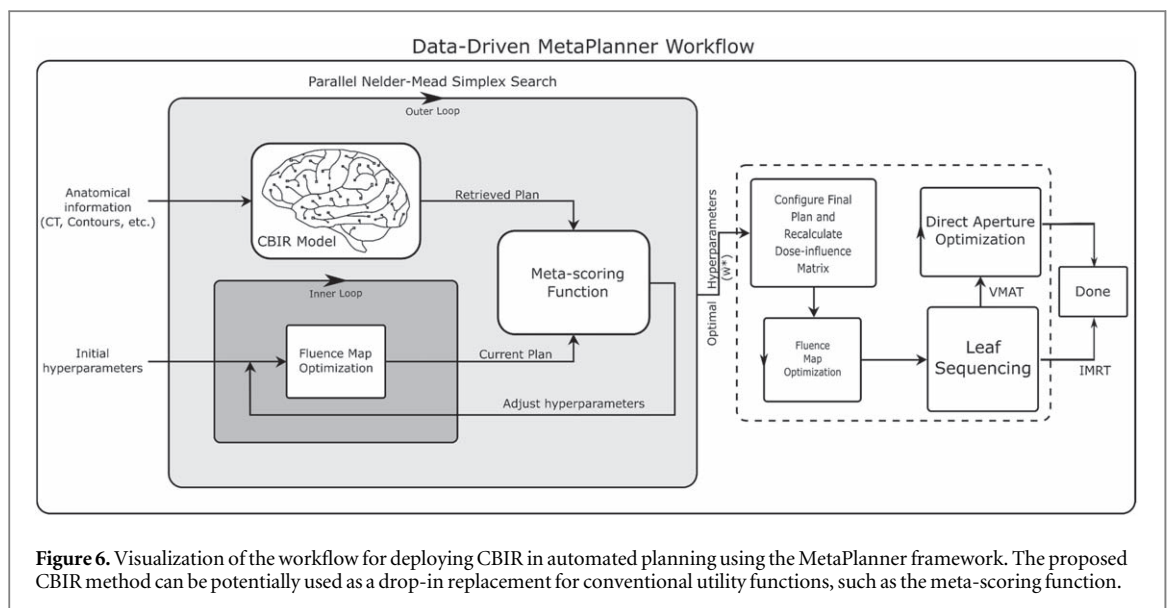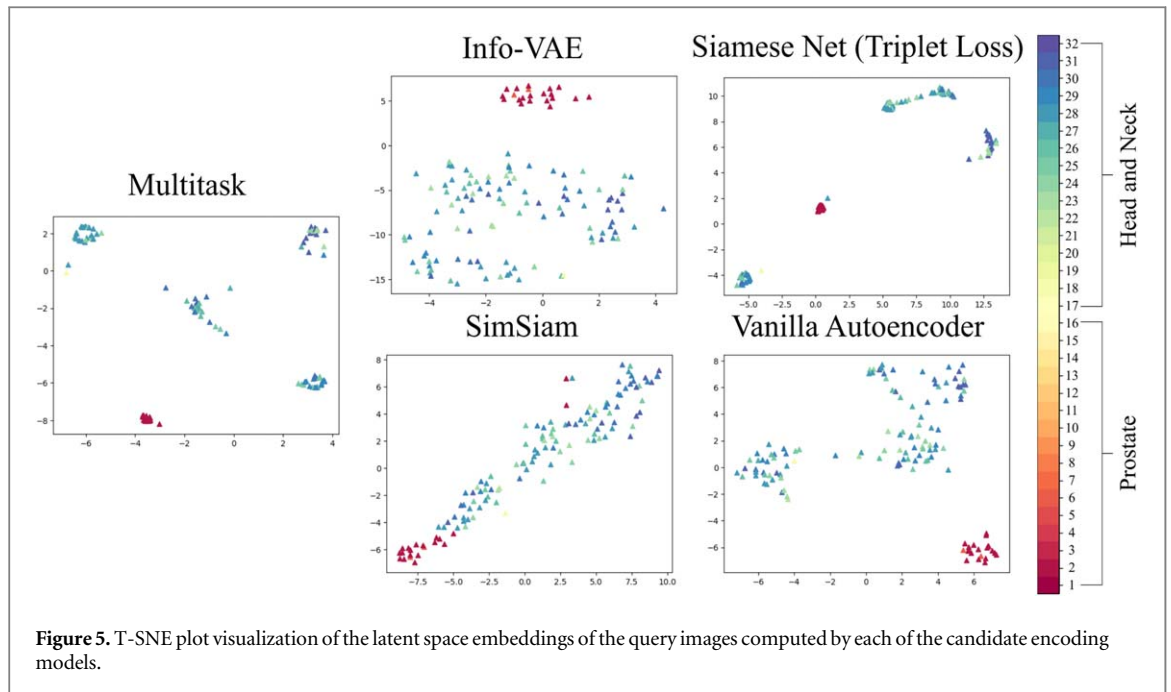
3.2. Clustering performance

We additionally evaluate clustering performance using metrics listed in the last 5 columns of table 2 (Hubert and Arabie 1985, Strehl and Ghosh 2003, Steinley 2004, Rosenberg and Hirschberg 2007). 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 $k = 1$. Of the benchmarked encoders, the top three performers for cluster homogeneity, cluster completeness, *V*-measure, adjusted Rand index, and adjusted mutual information were the MSN, SNTL, and Info VAE models (table 2). Figure 5 shows a TSNE plot of the latent space embeddings for the query set, with the evaluation class labels 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.

3.3. 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. 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 of the same classification as

8

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, where $k = 1$.

	Image retrieval performance				Clustering performance				
	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. T-SNE plot visualization of the latent space embeddings of the query images computed by each of the candidate encoding models.

Figure 6. Visualization of the workflow for deploying CBIR in automated planning using the MetaPlanner framework. The proposed CBIR method can be potentially used as a drop-in replacement for conventional utility functions, such as the meta-scoring function.

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 images relevant to treatment planning (i.e. CT, contours, dose distribution, etc) from a database. These retrieved images can be subsequently used in automated treatment planning pipelines to automate the iterative adjustments of optimization hyperparameters. An example of one such automated planning method is the MetaPlanner framework (Huang *et al* 2022), and we provide an example workflow incorporating those methods in figure 6. 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.

    The proposed CBIR framework will retrieve treatment plans from a database and can be utilized in any pipeline that would otherwise incorporate end-to-end knowledge-based planning. Specifically, the retrieved

dose distributions can be used in methods which directly optimize machine parameters through dose mimicking (McIntosh *et al* 2017, Mahmood *et al* 2018, Eriksson and Zhang 2022). 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 CBIR options. 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 (Lecun *et al* 1998, Deng *et al* 2009). Given the relatively small dataset size used in the 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.

In future work, we plan to incorporate the proposed image retrieval method into automated planning workflow. To clarify the various components for such a process, we describe an example implementation using image retrieval to create a data-driven utility function for automated planning (figure 6). In this example, a CBIR system is deployed to retrieve a reference plan (i.e. dose distribution, DVHs, etc) from the database based on similarity to the query patient. We can then compute a distance metric (utility function) between the reference plan and the query patient's plan that is undergoing optimization in order to guide automated planning. Currently, many automated planning methods model the treatment planning process as one where planners navigate the trade-offs of Pareto optimal solutions using a hand-crafted utility function. However, incorporating the retrieved dose distributions from an image retrieval approach enables the use of data-driven utility functions for automated planning, potentially providing a better model of the decision-making process.

This study is additionally 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, Yuan *et al* 2019, Dong and Xing 2020, Nomura *et al* 2020, Chen *et al* 2020b, Liang *et al* 2021, 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 the current study by evaluating alternative 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 knowledge-based planning 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 with regards to retrieval performance, clustering performance, and visual quality.

Acknowledgments

Data availability statement

The data cannot be made publicly available upon publication because they contain sensitive personal information. The data that support the findings of this study are available upon reasonable request from the authors.

Ethical statement

This research was conducted in accordance with the principles embodied in the Declaration of Helsinki and in accordance with local statutory requirements. The dataset used in this study consists of both public and institutional data. Approval and consent was obtained (Stanford IRB protocol #41335) for the collection, de-identification, and analysis of any institutional data used in this study.

ORCID iDs

Charles Huang ⑩ https://orcid.org/0000-0001-7559-8316
Varun Vasudevan ⑩ https://orcid.org/0000-0001-5013-6836
Md Tauhidul Islam ⑩ https://orcid.org/0000-0001-6259-632X
Yusuke Nomura ⑩ https://orcid.org/0000-0002-2180-568X

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** 977

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

Caruana R 1997 Multitask learning *Mach. Learn.* **28** 41–75

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 arXiv:2002.05709

Chen X and He K 2020 Exploring simple siamese representation learning arXiv:2011.10566

Chen Y, Xing L, Yu L, Bagshaw H P, Buyyounouski 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 multiobjective 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 Conf. on Computer Vision and Pattern Recognition 2009 IEEE Conf. 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.* **32** 2687–704

Eriksson O and Zhang T 2022 Robust automated radiation therapy treatment planning using scenario-specific dose prediction and robust dose mimicking *Med. Phys.* **49** 3564–73

Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press)

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 (Cambridge, MA: MIT Press) Online: https://papers.nips.c./paper/2006/hash/e9fb2eda3d9c55a0d89c98d6c54b5b3e-Abstract.html

Grill J-B *et al* 2020 Bootstrap your own latent: a new approach to self-supervised Learning arXiv:2006.07733

Huang C, Nomura Y, Yang Y and Xing L 2022 Meta-optimization for fully automated radiation therapy treatment planning *Phys. Med. Biol.* **67** 055011

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.* **68** 2907–17

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

Kingma D P and Welling M 2013 Auto-encoding variational bayes arXiv:1312.6114

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 Phys. Stat arXiv: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

Mogotsi 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

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. Biol.* **65** 215007

Pastor-Serrano O and Perkó Z 2021 Learning the Physics of Particle Transport via Transformers arXiv: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 *Proc. of the 2007 Joint Conf. 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 (Boston, MA, 07-12 June 2015)* (Picastaway, NJ: IEEE) pp 815–23

Sethi A 2018 *Khan's Treatment Planning in Radiation Oncology* 4th Edn, ed Faiz M Khan *et al* (Philadelphia, PA: Lippincott Williams & Wilkins (Wolters Kluwer)) vol 2016, p 648 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.* **47** 2329–36

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 Int. Symp. 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 arXiv: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