Medical Image Analysis 97 (2024) 103276

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report

Evi M.C. Huijben [1,a,b], Maarten L. Terpstra [2,3,a,b], Arthur Jr. Galapon [4,b], Suraj Pai [5,b], Adrian Thummerer [4,6,b], Peter Koopmans [7,b], Manya Afonso [8,b], Maureen van Eijnatten [1,b], Oliver Gurney-Champion [9,10,b], Zeli Chen [11], Yiwen Zhang [11], Kaiyi Zheng [11], Chuanpu Li [11], Haowen Pang [12], Chuyang Ye [12], Runqi Wang [13], Tao Song [14], Fuxin Fan [15], Jingna Qiu [15], Yixing Huang [15], Juhyung Ha [16], Jong Sung Park [16], Alexandra Alain-Beaudoin [17], Silvain Bériault [17], Pengxin Yu [18], Hongbin Guo [19], Zhanyao Huang [19], Gengwan Li [20], Xueru Zhang [20], Yubo Fan [21], Han Liu [21], Bowen Xin [22], Aaron Nicolson [22], Lujia Zhong [23], Zhiwei Deng [23], Gustav Müller-Franzes [24], Firas Khader [24], Xia Li [25], Ye Zhang [25], Cédric Hémon [26], Valentin Boussot [26], Zhihao Zhang [27], Long Wang [27], Lu Bai [28], Shaobin Wang [28], Derk Mus [29], Bram Kooiman [29], Chelsea A.H. Sargeant [30], Edward G.A. Henderson [30], Satoshi Kondo [31], Satoshi Kasai [32], Reza Karimzadeh [33], Bulat Ibragimov [33], Thomas Helfer [34], Jessica Dafflon [35,36], Zijie Chen [37], Enpei Wang [37], Zoltan Perko [38,b], Matteo Maspero [2,3,b,*]

[1] *Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands*
[2] *Radiotherapy Department, University Medical Center Utrecht, Utrecht, The Netherlands*
[3] *Computational Imaging Group for MR Diagnostics & Therapy, University Medical Center Utrecht, Utrecht, The Netherlands*
[4] *Department of Radiation Oncology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*
[5] *Department of Radiation Oncology (Maastro), GROW School for Oncology, Maastricht University Medical Centre, Maastricht, The Netherlands*
[6] *Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany*
[7] *Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands*
[8] *Wageningen University & Research, Wageningen Plant Research, Wageningen, The Netherlands*
[9] *Department of Radiology and Nuclear Medicine, Amsterdam UMC, location University of Amsterdam, Amsterdam, The Netherlands*
[10] *Cancer Center Amsterdam, Imaging and Biomarkers, Amsterdam, The Netherlands*
[11] *School of Biomedical Engineering, Southern Medical University, Guangzhou, China*
[12] *School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China*
[13] *School of Biomedical Engineering, ShanghaiTech University, Shanghai, China*
[14] *Fudan University, Shanghai, China*
[15] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[16] *Indiana University, Bloomington, USA*
[17] *Advanced Development Engineering, Elekta Ltd, Montreal, Canada*
[18] *Infervision Medical Technology Co., Ltd. Beijing, China*
[19] *Department of Biomedical Engineering, Shantou University, China*
[20] *Independent researchers*
[21] *Department of Computer Science, Vanderbilt University, Nashville, USA*
[22] *Australian e-Health Research Centre, CSIRO, Herston, Queensland, Australia*
[23] *Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California (USC), Los Angeles, CA, USA*
[24] *University Hospital Aachen, Aachen, Germany*
[25] *Center for Proton Therapy, Paul Scherrer Institut, Villigen, Switzerland; Department of Computer Science, ETH Zurich, Zurich, Switzerland*
[26] *University Rennes 1, CLCC Eugène Marquis, INSERM, LTSI, Rennes, France*
[27] *Subtle Medical, Shanghai, China*
[28] *MedMind Technology Co. Ltd., Beijing, China*
[29] *MRI Guidance BV, Utrecht, The Netherlands*
[30] *Division of Cancer Sciences, The University of Manchester, United Kingdom*
[31] *Muroran Institute of Technology, Hokkaido, Japan*

* Correspondence to: UMC Utrecht, Heidelberglaan 100, 3508 GA, P.O. Box 85500, Utrecht, The Netherlands.
E-mail address: m.maspero@umcutrecht.nl (M. Maspero).
[a] Equally contributing first authors.
[b] Challenge Organizer.

https://doi.org/10.1016/j.media.2024.103276
Received 13 March 2024; Received in revised form 2 June 2024; Accepted 11 July 2024
Available online 17 July 2024

[32] Niigata University of Health and Welfare, Niigata, Japan
[33] Image Analysis, Computational Modelling and Geometry, University of Copenhagen, Denmark
[34] IACS, Stony Brook University, NY, USA
[35] Data Science and Sharing Team, Functional Magnetic Resonance Imaging Facility, National Institute of Mental Health, Bethesda, USA
[36] Machine Learning Team, Functional Magnetic Resonance Imaging Facility National Institute of Mental Health, Bethesda, USA
[37] Shenying Medical Technology (Shenzhen) Co., Ltd., Shenzhen, Guangdong, China
[38] Delft University of Technology, Faculty of Applied Sciences, Department of Radiation Science and Technology, Delft, The Netherlands

ARTICLE INFO

Keywords:
Synthetic CT generation
Radiotherapy
Deep learning
Medical image synthesis

ABSTRACT

Radiation therapy plays a crucial role in cancer treatment, necessitating precise delivery of radiation to tumors while sparing healthy tissues over multiple days. Computed tomography (CT) is integral for treatment planning, offering electron density data crucial for accurate dose calculations. However, accurately representing patient anatomy is challenging, especially in adaptive radiotherapy, where CT is not acquired daily. Magnetic resonance imaging (MRI) provides superior soft-tissue contrast. Still, it lacks electron density information, while cone beam CT (CBCT) lacks direct electron density calibration and is mainly used for patient positioning.

Adopting MRI-only or CBCT-based adaptive radiotherapy eliminates the need for CT planning but presents challenges. Synthetic CT (sCT) generation techniques aim to address these challenges by using image synthesis to bridge the gap between MRI, CBCT, and CT. The SynthRAD2023 challenge was organized to compare synthetic CT generation methods using multi-center ground truth data from 1080 patients, divided into two tasks: (1) MRI-to-CT and (2) CBCT-to-CT. The evaluation included image similarity and dose-based metrics from proton and photon plans.

The challenge attracted significant participation, with 617 registrations and 22/17 valid submissions for tasks 1/2. Top-performing teams achieved high structural similarity indices ($\geq$ 0.87/0.90) and gamma pass rates for photon ($\geq$ 98.1%/99.0%) and proton ($\geq$ 97.3%/97.0%) plans. However, no significant correlation was found between image similarity metrics and dose accuracy, emphasizing the need for dose evaluation when assessing the clinical applicability of sCT.

SynthRAD2023 facilitated the investigation and benchmarking of sCT generation techniques, providing insights for developing MRI-only and CBCT-based adaptive radiotherapy. It showcased the growing capacity of deep learning to produce high-quality sCT, reducing reliance on conventional CT for treatment planning.

1. Introduction

More than half of cancer patients receive radiotherapy as the standard care, providing effective local treatment (Chandra et al., 2021). Radiotherapy is typically delivered daily over several weeks (Mitchell, 2013), aiming to provide a high radiation dose to the target while minimizing the dose to the surrounding healthy tissue. To achieve conformal radiation treatment, obtaining an electron density map of the patient's anatomy is crucial to determine beam attenuation and local dose deposition (Grégoire and Mackie, 2011). This electron density information is currently obtained through computed tomography (CT) (Seco and Evans, 2006). However, tumors are not always clearly visible on CT, and magnetic resonance imaging (MRI) has been proposed as its superior soft-tissue contrast offers improved visibility of tumor-boundaries and organs-at-risk (OARs) (Schmidt and Payne, 2015). Moreover, throughout the treatment course, patient anatomy may vary. In adaptive radiotherapy, new treatment plans are generated weekly or daily while the patient is on the treatment couch to maintain dose conformality. During adaptive radiotherapy, typically cone-beam CT (CBCT) (Nijkamp et al., 2008) or MRI (Lagendijk et al., 2014) are the sole imaging modalities at hand. However, neither MRI nor CBCT allows for direct treatment plan optimization as accurate electron density information is lacking. Techniques have been developed to generate synthetic CT (sCT) (also called pseudo-CT, virtual CT, surrogate CT) from MRI and CBCT to aid in determining local beam attenuation and dose deposition for treatment planning (Edmund and Nyholm, 2017). The sCT generation has paved the way for MRI-based treatment planning (MRI-only radiotherapy) and CBCT-based adaptive radiotherapy, which avoid additional radiation exposure due to imaging and reduce the treatment centers' workload by omitting unnecessary scans.

Although several approaches for obtaining sCT exist, including bulk density override and atlas-based methods, deep neural networks have recently shown promise in generating sCT (Spadea et al., 2021). Neural networks can be broadly categorized into convolutional neural networks (CNNs), e.g., U-net (Ronneberger et al., 2015), generative adversarial networks (GANs), e.g., cycleGAN, pix2pix, Goodfellow et al. (2014), Zhu et al. (2017), Isola et al. (2017), and, more recently, (vision-)transformers (Vaswani et al., 2017; Dosovitskiy et al., 2020) and diffusion models (Ho et al., 2020). Paired (supervised) and unpaired (unsupervised) training approaches have been suggested depending on the network architecture. The models were trained using 2-dimensional (2D) slices or 3D CT and MRI/CBCT volumes. Moreover, 2.5D approaches considering neighboring slices or perpendicular planes have been introduced to deal with spatial information and coherence while maintaining performance and feasible memory use. Most of these papers claim that their sCT generation method outperforms others. However, networks are often trained on different datasets and anatomies and evaluated using different metrics, making consistent methodological comparison difficult. Moreover, most sCT methods are evaluated based on image similarity metrics, whereas what matters is, ultimately, the effect of sCT on the treatment plan dose distribution, and image metrics do not necessarily reflect the dose accuracy (Kieselmann et al., 2018). This lack of a fair comparison hinders the identification of the best network design choices that should be implemented in clinical sCT tools.

To address these issues and provide a fair comparison, we organized the SynthRAD2023 Grand Challenge, held in conjunction with MICCAI 2023. In the challenge, we provided ground truth data and developed methods to facilitate fair model comparisons and increase the understanding of how different network designs influence performance. This challenge encourages the development and evaluation of state-of-the-art algorithms for generating accurate and clinically relevant sCT images from MRI and CBCT data. Two tasks were defined based on a new publicly available dataset (Thummerer et al., 2023a): (1) MRI-to-CT generation for MRI-only radiotherapy and MRI-guided radiotherapy and (2) CBCT-to-CT generation for image-guided adaptive radiotherapy (IGART) and online adaptive radiotherapy.

This paper reviews the challenge participation, evaluation, and ranking of the submitted algorithms based on image similarity and dose assessment for sCTs compared to ground truth CTs. The analysis

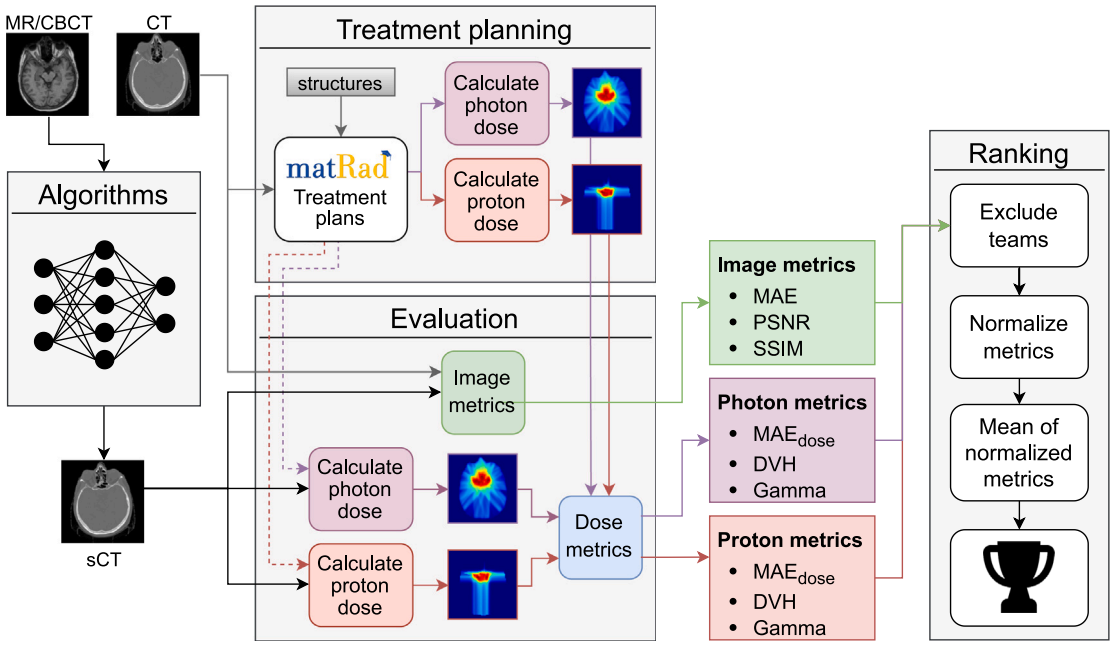


Fig. 1. The SynthRAD2023 pipeline. Left: the participants' algorithms generate sCT from input MRI or CBCT images. Middle block: the obtained sCT is evaluated with image similarity metrics (comparing sCT images to ground truth CT images) and dose metrics (comparing dose distributions recalculated on sCT and ground truth CT for pre-planned photon and proton treatment plans). Right: after calculating the metrics, the winner is determined by applying a ranking approach.

explores trends in submitted algorithms and their correlation with overall performance, focusing on the impact of variation within the dataset, the metrics chosen for evaluation, and examining ranking stability.

2. Material and methods

2.1. Challenge setup

The SynthRAD (Synthesizing Computed Tomography for Radiotherapy) challenge allowed teams to test and compare their sCT algorithms. The challenge was hosted on the Grand Challenge website https://synthrad2023.grand-challenge.org. It consisted of two tasks: task 1 involved generating sCT from MRI data, while task 2 focused on developing sCT from CBCT data. Each task comprises two subtasks involving the brain and pelvis regions.

The organizing team arose from the "Image synthesis & reconstruction" expertise subgroup of the Dutch deep learning in radiotherapy initiative www.DLinRT.org. The organizing group encompasses early-stage researchers, PhDs, postdocs, four assistant professors, and one associate professor from five Dutch University Medical Centers and three Dutch Technical Universities.

Fig. 1 presents an overview of the SynthRAD2023 Grand Challenge design, including the algorithms developed by the participants and the evaluation and ranking procedures performed by the organizers. Participants in the challenge were tasked with developing and training models capable of generating accurate sCT images using only input MRI or CBCT. Participants could participate in either task 1, task 2, or both. Only fully automated methods trained from scratch on the provided data could be used; in other words, pre-trained models were not allowed. The submissions were automatically evaluated on the Grand Challenge environment. Further details regarding participation rules and policies can be found in the Appendix A. As the sCTs are intended for radiotherapy, we analyzed photon and proton dose metrics alongside image similarity metrics, as described in Section 2.5. To determine the winner of the challenge, we ranked the teams based on these metrics, for which we provide a further explanation in Section 2.6. To ensure transparency and enable further exploration of the methods employed during the challenge, the data preprocessing and evaluation code can be accessed at https://github.com/SynthRAD2023.

2.2. Challenge phases

The challenge was divided into four phases: training, validation, preliminary test, and test. Teams had two months to familiarize themselves with the challenge and begin training their algorithms, as the training data was released on April 1, 2023. The validation phase began on June 1, 2023, and was a Type-1 challenge in which participants were required to execute the inference locally and submit the corresponding sCTs. This phase allowed for up to two submissions every four days, and the submitted sCTs were automatically assessed using image similarity metrics. The results were then updated on an open leaderboard, allowing real-time comparison between participating teams. The ground truth CT images used for validation were not shared with the participants to prevent biased results. The final test phase was a Type-2 challenge in which teams had to upload a Docker image containing their method, which is inferred and evaluated on the Grand Challenge platform. The test data and ground truth CTs were kept hidden. To familiarize participants with a Type-2 challenge, we introduced the preliminary test phase, which started on May 1, 2023. The preliminary test phase used six cases; only image similarity metrics were evaluated. The final test phase started on July 16, 2023, and lasted five weeks. The preliminary test phase and test phase ended on August 22, 2023. Teams were required to upload a Docker image of their algorithm and a description of their methods. To minimize algorithm tweaking to the test data, each team could submit only twice during the testing phase, and only the last submission was counted. The second submission allowed participants to correct potential errors arising during the first submission. During this phase, the generated sCT images underwent an image similarity evaluation and a photon and proton dose evaluation to verify the most relevant metrics for radiotherapy. The image similarity metrics were calculated online on the platform provided by Grand Challenge, and the dose evaluation was performed offline due to the computational resources required. At the end of the testing phase, the final ranking was published to show the performance of the participating teams. After the challenge, a post-challenge test phase was opened, and the preliminary and validation phases were reopened to enable continuous evaluation of algorithms until September 20, 2028.

2.3. Dataset

Data from 1080 patients undergoing radiotherapy treatment were included in the SynthRAD2023 dataset. The dataset consisted of imaging data from three Dutch University Medical Centers. Both task 1 (MRI-to-CT) and task 2 (CBCT-to-CT) included data from 270 patients for both the brain and pelvis anatomy (leading to $2 \times 2 \times 270$ image pairs). The 270 cases were divided into a training, validation, and test set of 180, 30, and 60 patients. The dataset consisted primarily of adult patients, with no gender restrictions applied. Only patients for whom the MRI or CBCT was acquired within two months of the CT were included to limit anatomical changes. It should be emphasized that the datasets for task 1 and task 2 did not contain the same patients. A detailed dataset description can be found in the publication by Thummerer et al. (2023a). Ethical approval was obtained from the data-providing institutes' internal review boards/Medical Ethical committees. The data was released under the CC BY-NC (Creative Commons Attribution-Non-Commercial) license and made available via Zenodo at https://zenodo.org/doi/10.5281/zenodo.7835406 (train), https://zenodo.org/doi/10.5281/zenodo.7868168 (validation), and https://doi.org/10.5281/zenodo.10514185 (test, available from 01-01-2028).

The imaging protocols used to acquire the MRI and CBCT adhered to the clinical routines of the individual centers. As a result, variations in the MRI, CBCT, and CT imaging protocols were present between centers and between datasets, which are representative of real-world application scenarios. A comprehensive table detailing the imaging parameters was provided alongside the dataset (Thummerer et al., 2023a). For task 1, MRIs were acquired with scanners from two different vendors using different settings per site. Additionally, centers A and C used MRI scanners with field strengths of 1.5T and 3T, while center B exclusively utilized a 1.5T scanner. T1-weighted gradient echo was selected for all brain data. The datasets from centers B and C included T1-weighted MRI acquired after Gadolinium contrast agent injection, whereas those from center A were acquired without contrast agent injection. The pelvis data comprised two-thirds of a T1-weighted gradient echo sequence and a T2-weighted spin echo sequence. For task 2, CBCTs were acquired with Linacs from two different vendors. The two sites that scanned CBCTs with Linacs from the same vendor had different acquisition protocols.

As described by Thummerer et al. (2023a), the data was preprocessed by resampling the voxel size to $1 \times 1 \times 1$ mm^3 for the brain and $1 \times 1 \times 2.5$ mm^3 for the pelvis patients, respectively. The face was intentionally removed for brain cases to protect patient privacy or proprietary information. The patient outline was automatically segmented on the MRI/CBCT using thresholding and morphological operations. This was followed by a dilation of 20 voxels in the axial plane and 2 in the superior-inferior directions. To ensure alignment between the MRI/CBCT and the CT, the field of view of the MRI/CBCT and CT was adjusted based on the patient outline, and rigid registration was performed. The resulting mask, including surrounding air, was provided and could be used by the participants for preprocessing.

2.4. Baseline algorithms

Two bulk-assignment baseline sCT models were used to provide insight into the evaluation metrics: "water" and "stratified". The water approach assigned 0 HU to voxels within the dilated body contour mask and −1000 HU outside the mask (air). As suggested by Maspero et al. (2017), a stratified approach was employed to obtain images resembling bulk-assigned sCT without geometrical deformations by starting from ground truth CT. Stratified sCTs were obtained by classifying the ground truth CT data into five categories and assigning bulk density values for voxels within their specific HU ranges. Voxels were categorized into five classes based on HU intensity levels, mapping a range of density values to a population-derived HU value for this tissue (Maspero et al., 2017), indicated as $\langle$lower bound, upper bound$\rangle$ HU → xx HU: 'air' ($\langle -\infty, -210\rangle$ HU → −968 HU), 'adipose tissue' ($[-210, -20\rangle$ HU → −86 HU), 'soft tissue' ($[-20, 120\rangle$ HU → 42 HU), 'bone marrow' ($[120, 555\rangle$ HU → 198 HU), and 'cortical bone' ($[555, \infty\rangle$ HU → 949 HU). The accuracy of the bone segmentation was further refined using a binary hole-filling algorithm to avoid soft tissue and air voxels within bone structures.

2.5. Evaluation

The sCTs generated by the participants were compared to the ground truth CTs based on metrics comparing image similarity and dose accuracy.

2.5.1. Image similarity

During the validation and test phases of the SynthRAD2023 Grand Challenge, the accuracy of the generated sCT images was evaluated using image similarity metrics within the dilated body contour masks $\mathcal{B} = \{i \mid \mathcal{M}_i = 1\}$ provided with the dataset. This evaluation aimed to assess how closely the sCTs resembled the reference CTs. The mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM) were considered as image similarity metrics, as they are commonly used in medical image synthesis (Spadea et al., 2021).

Masked MAE was calculated to measure the average absolute difference between corresponding voxels in the sCT and CT, defined as

$$\mathrm{MAE}(\mathrm{CT}, \mathrm{sCT}) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left|\mathrm{CT}_i - \mathrm{sCT}_i\right| \tag{1}$$

in which we sum over the voxels inside the body contour $\mathcal{B}$ and normalized by the total number of masked voxels $|\mathcal{B}|$.

Masked PSNR was calculated to quantify the ratio of maximum signal intensity over the noise level in the sCT compared to the CT, defined as

$$\mathrm{PSNR}(\mathrm{CT}, \mathrm{sCT}) = 10 \log_{10} \left(\frac{Q^2}{\frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(\mathrm{CT}_i - \mathrm{sCT}_i\right)^2} \right), \tag{2}$$

where Q is the dynamic range of the voxel intensities ($[-1024, 3000]$ HU). The CT and sCT were clipped to the dynamic range to calculate the masked PSNR.

Masked SSIM was calculated to assess structural similarity between CT and sCT. The SSIM for a voxel i between two images x and y is computed by

$$\mathrm{SSIM}_i(x, y) = \frac{\left(2\mu_x^i \mu_y^i + c_1\right)\left(2\sigma_{xy}^i + c_2\right)}{\left(\left(\mu_x^i\right)^2 + \left(\mu_y^i\right)^2 + c_1\right)\left(\left(\sigma_x^i\right)^2 + \left(\sigma_y^i\right)^2 + c_2\right)}, \tag{3}$$

where μ_x^i and σ_x^i are the mean and variance, respectively, of x within an $N \times N \times N$ window centered on voxel i and σ_{xy}^i is the covariance of x and y within an $N \times N \times N$ window centered on voxel i. $N = 7$ is the window size, and $c_1 = (0.01 \cdot L)^2$ and $c_2 = (0.03 \cdot L)^2$ are normalization constants, where $L = (3000 - (-1024))$ HU is the dynamic range of the volumes. The final masked SSIM value is then obtained by computing

$$\mathrm{SSIM}(\mathrm{CT}, \mathrm{sCT}) = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \mathrm{SSIM}_i(\mathrm{CT}, \mathrm{sCT}), \tag{4}$$

where the intensities of both the CT and sCT were clipped to $[-1024, 3000]$ HU and then adjusted to be non-negative by adding 1024 HU.

2.5.2. Dose distribution similarity

Photon and proton intensity-modulated treatment plans were optimized based on the reference CT using the matRad treatment planning system (Wieser et al., 2017). The dose was prescribed to the planning target volume (PTV) for simplicity in both modalities, i.e., no robust optimization was performed for proton plans, with specific doses and

Table 1
Dose constraints and planning objectives used in matRad for the brain and pelvis cases, respectively.
Source: Values are taken from Lambrecht et al. (2018) and Hall et al. (2021).

Brain 30 × 2.0 Gy to 95% of the PTV		Pelvis 20 × 3.0 Gy to 95% of the PTV	
Structure	**Constraint**	**Structure**	**Constraint**
Brainstem	$D_{0.03\,cc} < 60\,Gy$ $D_{0.03\,cc} < 54\,Gy$	Rectum	$V_{60\,Gy} < 1\%$ $V_{50\,Gy} < 22\%$ $V_{40\,Gy} < 38\%$ $V_{30\,Gy} < 57\%$ $V_{20\,Gy} < 85\%$
Chiasm	$D_{0.03\,cc} < 55\,Gy$	Bladder	$V_{60\,Gy} < 3\%$ $V_{56.8\,Gy} < 5\%$ $V_{48\,Gy} < 25\%$ $V_{40\,Gy} < 50\%$
Optical Nerve	$D_{0.03\,cc} < 55\,Gy$	Femur heads	$D_{max} < 37\,Gy$
Cochlea	$D_{mean} < 45\,Gy$	Colon	$D_{max} < 50\,Gy$
Brain	$V_{60\,Gy} < 3\,cc$	Small bowel	$D_{max} < 40\,Gy$

isodose levels for the brain and pelvis. Only co-planar plans were considered, with photon plans utilizing 9–13 equi-angled 6 MV beams from a generic Linac model and proton plans utilizing 3–4 beams (from bilateral and opaque angles) from a generic proton system available in matRad. To reduce the dose to the healthy tissues and to ensure plan uniformity between patients, we used the same objective functions and constraints available in matRad per treatment site. OAR dose limits were treated as hard constraints whenever possible and were revised on a patient-specific basis when hard constraints were not achievable. For a few patients, the number of beams and some optimization parameters (e.g., optimizer, maximum number of iterations, and objective weights) were also fine-tuned to meet dose prescriptions and OAR limits. All planning goals and OAR dose limits were based on international guidelines for the brain (Lambrecht et al., 2018) and pelvis (Hall et al., 2021) are summarized in Table 1.

Throughout the dose evaluation process, the dose was recalculated on each sCT for both proton and photon treatment plans. This recalculation was carried out without propagating organ delineations or replanning, a deliberate measure taken to avoid potential differences arising from plan optimization. Subsequently, the differences between the planning dose distributions, originally calculated on CT, and the recalculated dose distributions on the sCT for both photon and proton plans were quantified using three specific metrics. To ensure high reproducibility and facilitate fair comparisons for the SynthRAD2023 test set, the offline dose evaluation will be available at https://doi.org/10.5281/zenodo.10514185 at the time of the release of the test set.

Relative mean absolute dose difference within high dose regions $\mathcal{H} = \{i \mid D_{CT,i} \geq 0.9 \cdot D_{prescribed}\}$ were calculated to assess the difference in received dose in and around the target, defined as

$$\mathrm{MAE}_{\mathrm{dose}} = \frac{1}{|\mathcal{H}|} \sum_{i \in \mathcal{H}} \frac{D_{CT,i} - D_{sCT,i}}{D_{\mathrm{prescribed}}}, \tag{5}$$

with $D_{(s)CT}$ being the dose distribution in the $(s)CT$ and $D_{prescribed}$ the prescribed dose.

Dose-volume histogram (DVH) parameters were calculated to assess the differences in the doses received by the PTV and OARs: the near-minimum dose in the PTV $D98_{\mathrm{PTV}}$, the PTV volume receiving at least 95% of the prescribed dose $V95_{\mathrm{PTV}}$, the near-maximum dose of a given OAR $D2_{\mathrm{OAR}}$, and the mean dose received by a given OAR $D\mathrm{mean}_{\mathrm{OAR}}$. Specifically, the use of the near-minimum and near-maximum was suggested by ICRU83 (https://www.fnkv.cz/soubory/216/icru-83.pdf). We included the relative absolute differences for all parameters as defined by

$$D98_{\mathrm{PTV,CT}} = \frac{|D98_{\mathrm{PTV,CT}} - D98_{\mathrm{PTV,sCT}} + \epsilon|}{D98_{\mathrm{PTV,CT}} + \epsilon}, \tag{6}$$

$$V95_{\mathrm{PTV,CT}} = \frac{|V95_{\mathrm{PTV,CT}} - V95_{\mathrm{PTV,sCT}} + \epsilon|}{V95_{\mathrm{PTV,CT}} + \epsilon}, \tag{7}$$

$$D2_{\mathrm{OARs}} = \frac{1}{n_{\mathrm{OARs}}} \sum_{\mathrm{OAR}} \frac{|D2_{\mathrm{OAR,CT}} - D2_{\mathrm{OAR,sCT}} + \epsilon|}{D2_{\mathrm{OAR,CT}} + \epsilon}, \tag{8}$$

$$D\mathrm{mean}_{\mathrm{OARs}} = \frac{1}{n_{\mathrm{OARs}}} \sum_{\mathrm{OAR}} \frac{|D\mathrm{mean}_{\mathrm{OAR,CT}} - D\mathrm{mean}_{\mathrm{OAR,sCT}} + \epsilon|}{D\mathrm{mean}_{\mathrm{OAR,CT}} + \epsilon}, \tag{9}$$

where $\epsilon = 1e{-}12$ to avoid division by zero and n_{OARs} is the number of OARs. For each patient, we used the three OARs (if available) that had the highest average of $D5_{\mathrm{OAR}}$ and $D\mathrm{mean}_{\mathrm{OAR}}$ to analyze dose differences in organs close to the target. We summed the four terms to obtain one final value for the DVH metric.

Gamma pass rates were calculated to compare the 3D spatial dose distributions from the sCTs with the dose obtained from the CT. This calculation followed the 3D gamma pass rate approach described by Low et al. (1998) with a dose-difference criterion (ΔD) of 2% and a distance-to-agreement criterion (Δd) of 2 mm. The gamma pass rate at each position vector in the sCT was determined by comparing it with the CT dose. Gamma pass rates were evaluated within regions receiving doses ≥ 10% of the prescribed dose (Ezzell et al., 2009).

2.6. Eligibility and ranking

The nine metrics defined above were calculated for each test case and aggregated across all test cases for each participating team ($\mu \pm \sigma$). Teams were not considered in the ranking if their method did not outperform the water baseline for all three individual image similarity metrics. Moreover, the participants' method must complete the generation of a single sCT within 15 min on the Grand-Challenge platform, as described in the Appendix A.

Several methods exist for creating a ranking for a challenge with multiple metrics, including (1) calculating the mean over all metrics and ranking the aggregated scores (MeanThenRank), (2) calculating the median over all metrics and ranking the aggregated scores (MedianThenRank), (3) calculating the ranking for each metric and computing the mean of the aggregated ranks (RankThenMean), and (4) Calculating the ranking for each metric and computing the median of the aggregated ranks (RankThenMedian). Directly applying MeanThenRank and MedianThenRank to the nine metrics is inappropriate due to their lack of normalization and the differing orderings (ascending or descending). To fairly rank the submissions, each metric was normalized and scaled between zero (indicating the worst average team performance) and one (indicating the best average team performance). Subsequently, the normalized metrics are used to calculate the mean or median and rank the aggregated score.

In the context of the SynthRAD2023 challenge, the MeanThenRank approach was employed to determine the winners. This method should account for variations in team performance, enabling a fair evaluation considering the diverse clinically relevant aspects of image similarity, photon dose, and proton dose metrics. To analyze biases introduced by the ranking method, we also studied how the other ranking approaches would affect the outcome to assess ranking stability.

2.7. Analysis

2.7.1. Overall sCT performance

Besides computing the aggregated metrics per submission ($\mu \pm \sigma$), we analyzed the significance of one team outperforming another in terms of individual metrics. To do so, we used the Wilcoxon signed-rank test (Wilcoxon, 1945) with Holm's adjustment for multiple testing (Holm, 1979) for each metric separately, offering insights into the pairwise performance differences between teams. The significance level for this test is set at $\alpha = 0.05$. Additionally, we recorded the inference time of the participant's methods ($\mu \pm \sigma$) to synthesize the CT from the CBCT or MRI data on the Grand Challenge infrastructure.

2.7.2. Model design predictors

We evaluated the model design choices adopted by participating teams thoroughly, aiming to identify the impact of these choices on overall ranking and performance. Statistical significance of the differences in SSIM performance within each subtask is determined using the Mann–Whitney U-test (Mann and Whitney, 1947) ($\alpha = 0.01$), chosen for its suitability in comparing two independent samples that may not adhere to normal distribution. This test is particularly robust in the context of our analysis, providing reliable insights into the performance disparities associated with distinct design choices. To define predictors for sCT performance, we analyzed different design choices. We categorized them into five key aspects: (1) model and anatomy, (2) backbone architecture, (3) spatial configuration, (4) preprocessing, (5) data augmentation, and (6) postprocessing.

Model and anatomy. Teams used different strategies to handle brain and pelvis data. Some teams used one collective model trained on both brain and pelvis patients ("One model") or conditioned the collective model on the anatomical region ("One model, anatomy conditional"). In contrast, others trained the same model separately for the brain and pelvis subsets ("Two identical models"). Additionally, some teams used the same or a similar backbone architecture for both regions with distinctions in training parameters or network layers ("Two identical backbones, different training param." and "Two similar models", respectively). Others employed entirely different models for the two regions ("Two different models").

Backbone architecture. The base of a synthesis model involved using a CNN encoder–decoder model, which is a standard choice for image reconstruction and translation within deep learning. Teams also explored alternative architectures, such as GAN-based models that introduced a discriminator network and adversarial loss, transformer-based architectures that emphasized attention in the synthesis process, and diffusion model-based approaches that relied on an iterative diffusion process during inference. Moreover, some teams used an ensemble of multiple models to produce the final output.

Supervision. Each team reported the supervision approach adopted. Supervised (paired) training was guided by directly comparing predictions (sCT) to ground truth (CT) from the same cases. Unsupervised (unpaired) training was guided by introducing cycle-consistency as introduced by Zhu et al. (2017).

Spatial configuration. The implementation of sCT generation models varied in different spatial configurations. Opting for fully 3D models was possible, considering the entire image volume as input. However, fully 3D models were often restricted in use by available computing resources; therefore, many studies employed 3D patch-based approaches, 2.5D models considering multiple consecutive 2D slices, or a combination of orthogonal slices, full slice 2D models, or 2D patch-based models.

Preprocessing. A range of preprocessing techniques were used in the submitted algorithms, focusing on resizing and intensity normalization, necessary for stable and optimal model training. Resizing was used to achieve the desired voxel size, such as in the case of iso-resampling or the desired model input size. Intensity normalization was implemented linearly at the population level, at the patient level, or as standardization by ensuring well-distributed data based on a specific mean and standard deviation. Furthermore, some teams used intensity clipping to remove outliers, applied histogram matching or provided a specialized pipeline for the specific modality or anatomical region being processed.

Data augmentation. Various data augmentation techniques ensured a diverse training set, potentially making the models more robust to unseen cases in the test set. Teams introduced randomness through random crop or patch selection, flipping, rotation, blurring, noise addition, and intensity transformations like bias field or contrast adjustments. Random deformations, whether affine or elastic, were also applied to enhance the diversity of the training set.

Table 2
Details on the challenge participation. Participants without a team are displayed as a one-person team.

	Validation	Preliminary test	Test
Task 1	275 valid submissions (from 38 teams)	77 valid submissions (from 27 teams)	18 included, 4 excluded, and 2 failed teams
Task 2	207 valid submissions (from 25 teams)	36 valid submissions (from 15 teams)	14 included and 3 excluded teams

Postprocessing. Some teams that implemented patch-based models averaged overlapping patches at test time. The multiple outputs of ensembled methods could be combined into a single sCT. Additionally, specific postprocessing steps were implemented considering prior knowledge of the modality, such as noise and artifact removal. The inversion of original preprocessing steps, such as normalization and padding, was crucial in obtaining the final sCT with accurate dimensions and voxel representation in Hounsfield units (HU).

2.7.3. Data influence

By examining the teams' performances, we analyzed the test dataset to identify the characteristics and features of the samples correlating with synthesized image quality. The analysis compared image similarity and dose metrics ($\mu \pm \sigma$) averaged for each task, center, and anatomy within the test set. In addition, for task 1, the influence of MR acquisition protocol and magnetic field strength on performance was investigated. Statistical significance between the groups was established using the Mann–Whitney U-Test (Mann and Whitney, 1947) ($\alpha = 0.01$). Lastly, we extended our analysis to a patient level, allowing for detailed evaluation of low-performing patients.

2.7.4. Metric correlations

For clarification throughout the paper, we defined the term 'metric group' to refer to one of the three categories of evaluation metrics: image similarity metrics (MAE, PSNR, and SSIM), photon dose metrics (MAE_{dose}, DVH_{metric}, and γ), and proton dose metrics (MAE_{dose}, DVH_{metric}, and γ).

Our objective was to analyze the correlation within and between metric groups. To achieve this, we employed visual assessments to illustrate correlations within a metric group. We used the Spearman rank correlation coefficient ρ (Spearman, 1904) to quantify correlations between all metrics. This coefficient considers the ordinal relationship between the ranks of single test case performances, providing robustness against variations in the scale and direction of the values.

2.7.5. Ranking stability and correlations

We used Kendall's τ correlation coefficient (Kendall, 1938) between the approaches to analyze the effect of ranking approach choice. This coefficient quantifies the correlations between the ranking approaches outlined in Section 2.6, assessing the similarity in the relative ordering of elements across different rankings.

In addition, we investigated the stability of the final rankings at a patient level, as recommended by Wiesenfarth et al. (2021). This involved implementing bootstrapping to examine variations in the ranking positions of all teams. The ranking process was iteratively applied to 1000 bootstrap sets. Each bootstrapping set consisted of 120 randomly selected patients from the test set, with patients potentially being selected more than once. The MeanThenRank approach was employed to rank the teams by first normalizing the metric values based on the best and worst average performance of that metric per bootstrap sample.

3. Participation

The SynthRAD2023 Grand Challenge witnessed substantial participation from research teams worldwide, showcasing various techniques

and methodologies for sCT generation. By the end of the test phase, the training dataset had been downloaded 1797 times, and 617 researchers had registered for the challenge, forming 94 teams and 429 individual participants. Participation in the challenge phases decreased over time, resulting in 22 and 17 successful submissions in the test phase for tasks 1 and 2, respectively. Based on the criteria described in Section 2.6, 18 and 14 teams were included in the analysis for tasks 1 and 2, respectively (Table 2). Note that due to an unexpectedly large matrix size for one patient in the test set, the inference time limit was raised to accommodate the sCT generation. Nine of the included teams participated in both tasks, primarily utilizing the same or similar models for both tasks. Tables 3 and 4 show an overview of the proposed methods of all teams for tasks 1 and 2, respectively. More detailed descriptions of the methods implemented by the top five teams for both tasks are presented in Sections 3.1 to 3.7. Detailed method descriptions of all other teams can be found in supplementary document A.

3.1. SMU-MedVision (task 1 & 2)

SMU-MedVision employed a hybrid 3D patch-based CNN and transformer Unet network with multi-scale structure extraction and preservation (MSEP) for task 1 (Chen et al., 2023; Zhong et al., 2023). In the encoder, they employed channel and spatial-wise attention to extract spatial information, allowing for varying input sizes. Additionally, a residual dilated Swin transformer (RDSformer) was integrated into each skip connection of the UNet to enhance the preservation of structural information in cross-modal features (Liu et al., 2021). Two identical models were created for both anatomical regions, including the masked MAE and VGG19 perceptual loss (Johnson et al., 2016). Preprocessing involved Z-score normalization tailored to individual patient statistics and random horizontal and vertical flipping for data augmentation. At test time, overlapping patches were created by selecting every 80,000th voxel within the body mask as the central point of each patch. These overlapping patches were averaged to result in the full sCT. The model underwent training for 200 epochs using the Adam optimizer with a learning rate of 2e−4 and a poly decay scheduler. The final epoch used at test time was determined based on the best MAE in the sub-validation set created from the training set.

For task 2, SMU-MedVision implemented a 2.5D Unet++ (Zhou et al., 2018) with a ResNeXt101 backbone, with the loss function combining masked MAE loss, VGG19 perceptual loss (Johnson et al., 2016), and L2 regularization. The model was trained using brain and pelvis data and then fine-tuned per region. Preprocessing involved resizing, clipping, and linear normalization of the CT. Training data augmentation included shift scale rotations with horizontal and vertical flipping, while test data underwent augmentation via horizontal and vertical flipping. Slices of 5 × 384 × 384 voxels were used for collective pretraining, and the model input sizes for fine-tuning were 5 × 288 × 288 voxels for the brain and 5 × 416 × 416 voxels for the pelvis. Postprocessing included the inversion of test-time augmentations. The model was collectively trained for 40 epochs and then fine-tuned using 5-fold cross-validation for 50 for the brain and 40 epochs for the pelvis, respectively, and optimized using an AdamW optimizer with a stepped decay learning rate schedule. The final result was based on an ensemble of all five folds (with the best validation MAE) and a model trained on the completely provided training set (for the number of epochs mentioned above).

3.2. Jetta_Pang (task 1)

Jetta_Pang implemented two 3D patch-based nnU-Net (Isensee et al., 2021) models with an MSE loss for task 1: Model-Brain and Model-Pelvis. Preprocessing involved Z-score normalization for MRI and no normalization for CT images. No resizing, rescaling, or data augmentation was applied. Model input sizes were 64 × 128 × 224 voxels for brain patches and 112 × 160 × 128 voxels for pelvis patches. Inference utilized the nnU-Net's default sliding window with half-patch size overlap, and no postprocessing steps were applied since the sCT was presented in HU. The models were trained for 1000 epochs using an SGD optimizer with a Nesterov momentum of 0.99, an initial learning rate of 1e−2, and a polyLR scheduler.

3.3. GEneRaTion (task 2)

GEneRaTion employed a 2D restoration approach using a Swin transformer (Liang et al., 2021) combined with a pre-trained masked autoencoder (He et al., 2022) for task 2. The SwinV2 architecture (Liu et al., 2022) was enhanced by incorporating group propagation blocks (Yang et al., 2022). Depending on the training stage, the model included either an L1, MSE, or perceptual loss (Johnson et al., 2016). Two identical models were created for the brain and pelvis. (CB)CT was linearly normalized between [−1000, 3000] HU for the brain and [−1000, 2000] HU for the pelvis. In a self-supervised pretraining phase, an L1 loss and a learning rate of 1e−4 were applied to 8 × 8 random patches with at least 75% of the patch within the provided body mask. Random 90° rotations or horizontal or vertical flipping were part of this pretraining step. Subsequently, the models were fine-tuned for 100 epochs on axial slices randomly cropped to 160 × 160 voxels, utilizing three stages. These stages involved training with (1) L1 loss and a learning rate of 1e−4, (2) MSE loss with a learning rate of 2e−5, and (3) a perceptual loss with a learning rate of 1e−5. During test-time ensembling, the three sCTs were combined through a weighted average, with weights calculated by $\frac{sCT-\text{mean}(sCT)}{\max(sCT)}$, and preprocessing steps were restored.

3.4. FAYIU (task 1 & 2)

Team FAYIU implemented a patch-based 3D Swin UNETR (Hatamizadeh et al., 2021) in MONAI (Cardoso et al., 2022) for both tasks and regions separately. The Swin UNETR architecture, incorporating a vision transformer-based encoder and CNN-based decoder, enabled the processing of 3D patches. The models used a masked L1 loss. MRI inputs were normalized by dividing by 1000, while (CB)CT inputs were first made non-negative and subsequently divided by 2000. For training, 20 random patches of 32 × 96 × 96 voxels were selected per patient, and no other data augmentation techniques were applied. At inference time, patches overlapping by 28 × 72 × 72 voxels were selected, and overlapping regions were averaged in a weighted manner, with the weights for adjacent patches decreasing linearly as the overlap distance increased. Furthermore, the CT normalization procedure was reverted to result in an sCT in HU. The models were trained for 4000 epochs using the Adam optimizer and step-wise learning rate decay from 5e−4 to 5e−5.

3.5. iu_mia (task 1 & 2)

Team iu_mia employed a 3D patch-based ShuffleUNet (Chatterjee et al., 2021) model conditioned on the anatomical region for both tasks, with the L1 loss for both tasks. This model incorporates specialized 3D pixel unshuffling and shuffling modules to effectively handle the unique 3D aspects of medical imaging data. Z-score normalization was applied to the 3D MRI volumes, while (CB)CT volumes underwent linear scaling by ((CB)CT − 1024)/4024. They selected random patches measuring 96 × 96 × 96 voxels for training, and no other data augmentation techniques were applied. At test time, sCTs were generated from patches with a 62.5% overlap and averaging using Gaussian weighting (σ = 0.125), and the normalization process was inverted. The models were trained for 3000 epochs using the Adam optimizer with a linear learning rate scheduler initialized at 1e−3.

Table 3
Ranking and model details task 1 (MRI-to-CT synthesis). When a check is used, this step is applied to both MRI and CT and brain (br) and pelvis (pel); otherwise, it is specified by the subgroup. All distinctions listed in the first two rows are described in Section 2.7.2.

Rank	Team	Model & anatomy						Backbone arch.						Sup.		Spatial config.			Preprocessing									Data augmentation								Postprocess.	
		One model	One model, anatomy conditional	Two identical models	Two identical backbones, different training param.	Two similar models	Two different models	CNN encoder-decoder	GAN	Transformer	Diffusion model	Ensemble of multiple models		Supervised	Unsupervised	3D	3D patch-based	2.5D	2D	Iso-resampling / resizing	Clipping	Patient level linear normalization	Population level normalization	Standardization	Histogram matching	Other (e.g. N4 correction, smoothing, limb removal)		Random crop / patch	Flipping	Rotation	Blurring	Noise addition	Intensity transform (bias field, contrast / histogram adj.)	Deformation (affine or elastic)		Average overlapping patches	Noise / artifact removal	Revert normalization / padding
1	SMU-MedVision			✓						✓				✓			✓							✓					✓							✓		✓
2	Jetta_Pang				✓			✓						✓			✓							MR												✓		
3	FAYIU			✓						✓				✓			✓						✓					✓								✓		✓
4	Elekta			✓					✓			✓		✓				✓		✓	✓		✓					✓			MR			MR	✓	✓		✓
5	iu_mia		✓					✓						✓					✓	pel			CT	MR				✓								✓		✓
6	ShantouBME		✓					✓						✓					✓			MR	CT					✓										✓
7	FGH_365		✓	✓					✓			✓		✓			✓				✓	MR	CT					✓			✓	✓	✓	✓	✓		✓
8	USC-LONI				✓						✓			✓				✓		✓		MR	CT													✓	✓
9	UKA			✓				✓				✓		✓				✓			✓		CT						✓						✓		✓
10	PSICPT_AI4PT			✓				✓						✓				✓					CT	MR				✓	✓	✓							✓
11	SubtleCT			✓				✓						✓				✓			✓	MR	CT														✓
12	mriG			✓				✓						✓			✓					✓						✓	✓	✓			✓		✓		✓
13	KoalAI					✓			✓			✓		✓			✓						CT		MR	✓		✓		✓			✓	✓	✓		✓
14	Breizh-CT			✓					✓					✓			✓						✓												✓		✓
15	SKJP				✓			✓						✓				✓				MR			MR												
16	reza.karimzadeh			✓					✓	✓				✓			✓						✓					✓		✓			✓		✓		✓
17	thomashelfer						✓		pel		br			✓		br	pel						✓													✓	
18	X-MAN	✓							✓					✓			✓					MR						✓				✓			✓		✓

Table 4
Ranking and model details task 2 (CBCT-to-CT synthesis). When a check is used, this step is applied to both CBCT (CB) and CT and the brain and pelvis (pl.). Otherwise, it is specified by the subgroup. All distinctions listed in the first two rows are described in Section 2.7.2.

Rank	Team	Model & anatomy					Backbone arch.					Sup.		Spatial config.			Preprocessing								Data augmentation							Postproc.		
		One model	One model, anatomy conditional	Two identical models	Two identical backbones, different training param.	Two similar models	CNN encoder-decoder	GAN	Transformer	Diffusion model	Ensemble of multiple models	Supervised	Unsupervised	3D patch-based	2.5D	2D	2D patch-based	Iso-resampling / resizing	Clipping	Patient level linear normalization	Population level linear normalization	Histogram matching	Other (e.g. overflow correction, multi-windowing)	Random crop / patch / translation	Flipping	Rotation	Noise addition	Intensity transform (contrast / histogram adj.)	Deformation (affine or elastic)	Average overlapping patches	Noise / artifact / background removal	Revert normalization / padding
1	SMU-MedVision		✓		✓		✓				✓	✓			✓			✓	✓		✓				✓	✓						✓
2	GEneRaTion			✓					✓		✓	✓				✓					✓			✓	✓	✓						✓
3	iu_mia		✓				✓					✓		✓				pel			✓			✓						✓			✓
4	FAYIU			✓					✓			✓		✓							✓			✓						✓			✓
5	Pengxin Yu			✓			✓					✓		✓							✓			✓						✓			✓
6	FGZ Medical Research	✓								✓		✓				✓		✓	✓		✓												✓
7	KoalAI					✓		✓			✓	✓		✓							✓		pel	✓		✓		✓	✓	✓			✓
8	FGH_365		✓	✓				✓			✓	✓		✓					✓	CB	CT			✓	✓		✓	✓	✓	✓			✓
9	UKA			✓			✓				✓	✓			✓				✓		✓				✓					✓			✓
10	Breizh-CT			✓				✓				✓					✓				✓	CB								✓			✓
11	MedicalMind			✓				✓				✓				✓		✓						✓	✓	✓					✓	✓	✓
12	RRRocket_Lollies			✓				✓					✓			✓		✓	✓		✓		✓											✓
13	SKJP				✓		✓					✓			✓						CB													
14	X-MAN	✓						✓				✓		✓						CB				✓				✓			✓		✓

3.6. Elekta (task 1)

Team Elekta only participated in task 1, where they employed a 2.5D pix2pix (Isola et al., 2017) model using a ResUnet (Zhang et al., 2018) generator and a discriminator implemented similarly to the encoding part of the ResUnet. Spectral normalization (Miyato et al., 2018) was applied after each convolutional layer, and instance normalization replaced group normalization. Two identical models were created per anatomical region, using the least squares GAN loss (Mao et al., 2017) with L1 regularization (weight of 50). Linear scaling of MRI and CT intensities was conducted to fit within the range of [−1, +1], with source ranges determined by percentiles for MRI and fixed as [−1000, +2200] HU or [−1000, +3000] HU for CT. Two networks were trained for both regions, one covering the full CT intensity range and the other focusing on a narrower range. The training involved randomly selecting axial patches of $5 \times 192 \times 192$ and augmented using affine transformations, synthetic multiplicative bias fields, blurring, sharpening, gamma contrast adjustments, and linear intensity transformations for MRI. For CT, only affine transformations were applied. During inference, patches with $4 \times 96 \times 96$ voxels overlap were combined through weighted averaging, with higher weights assigned to pixels near the center of the patch and lower weights to those near the edge. The model was trained using the Adam optimizer and learning rates of 1e−4 and 5e−5 for the generator and discriminator, respectively. Additionally, a slow-moving exponential moving average (EMA) of the generator parameters was tracked during training and used as the final model for inference. Each model was trained six times, resulting in a final sCT ensembled by averaging the results of the six models.

3.7. Pengxin Yu (task 2)

Pengxin Yu employed a 3D patch-based model inspired by Ge et al. (2019) for task 2, implemented separately for the brain and pelvis. The model architecture featured consecutive multiscale residual blocks, effectively extracting fine-grained spatial structures and integrating stereo-correlation and image-expression constraints alongside the L1 reconstruction loss to guide structural detail and scene content. CBCT was linearly normalized between [−1000, 2000] HU, and for CT, center and region-specific windows were set: brain center A: [0, 3000] HU, brain center B and C: [−1000, 2000] HU, pelvis center A: [0, 2000] HU, and pelvis center B and C: [−1000, 1000] HU, after which the intensities were linearly normalized. During training, patches of $8 \times 180 \times 180$ voxels were created by randomly resizing, cropping, and horizontal flipping. At test time, overlapping patches were selected with an overlap of $2 \times 32 \times 48$ voxels. The models were trained for 1000 epochs with the AdamW optimizer with an initial learning rate of 3e−4 and reducing the learning rate by a factor 10 when the validation loss has not decreased for 10 epochs in a row. The final epoch used at test-time was determined based on the best PSNR on the sub-validation set, created from the training set.

4. Results

4.1. Overall sCT generation performance

Table 5 presents the final ranking and quantitative results of the 18 eligible teams for task 1 and 14 eligible teams for task 2, along with the two baseline algorithms. All eligible teams outperformed the water baseline in both tasks based on the image similarity metrics. Almost all teams also outperform the water baseline based on the dose metrics. However, one team (X-MAN) did not outperform the water baseline when considering the γ_{photon} and $\text{DVH}_{\text{proton}}$ metrics in task 1 and the $\text{MAE}_{\text{photon}}$ and γ_{photon} metrics in task 2. On the other hand, 11/18 and 14/18 teams outperform the stratified baseline based on image similarity for tasks 1 and 2, respectively. Regarding the dose metrics in task 1, 10/18 and 14/18 outperformed the stratified baseline for the photon and proton gamma pass rate, respectively. In task 2, the stratified baseline outperformed all teams based on the photon gamma pass rate, while 11/15 teams achieved a higher proton gamma pass rate than the stratified baseline. Interestingly, a higher image similarity did not automatically lead to an improved dose distribution. For example, comparing SMU-MedVision (rank 1) and FGZ Medical Research (rank 6) for task 2, we observe a large difference in MAE of 49.95 ± 11.78 and 60.65 ± 12.56 HU, while a subtle difference in photon gamma pass rates of 99.49 ± 1.65 and 99.57 ± 1.07 is seen. Such differences motivated us to perform an in-depth statistical analysis examining the significance of one team outperforming another based on individual metrics (Figures 1 and 2 in supplementary document B). Based on the image similarity metrics, high-ranking teams robustly outperform lower-ranked teams. Statistical significant improvements were observed when comparing all image metrics between a team and another team ranked at least seven places lower for task 1, or six placed lower for task 2. However, for the dose metrics, this relation is weaker. In task 1, no statistical significant differences were observed between the top fourteen teams regarding the photon dose metrics and top eleven teams regarding the proton dose metrics. In task 2, no statistically significant differences were observed between the top eight teams regarding the photon and proton dose metrics, except for the fifth team (Pengxin Yu), which significantly outperforms the seventh team (KoalAI) regarding the proton DVH metric.

Overall, the teams successfully generated high-quality sCTs, accurately synthesizing soft-tissue density. However, visual examples in Fig. 2 show more pronounced errors at transitions between tissue densities, such as the boundaries between air and soft tissue or soft tissue and bone. These errors at the boundaries of the input with the ground truth CT appear consistent across teams and lead to increased dose error when a beam passes through these regions. Moreover, in the pelvic cases, the anatomy does not always fit within the field-of-view of the CBCT, requiring participants to synthesize anatomy not present in the model input.

The average inference time per case was 5.2 ± 2.8 minutes, with teams utilizing an average of 4.0 ± 4.8 GB of GPU RAM. The maximum observed inference time for a single case was 21.8 min. There was a notable spread in resource usage between teams, and a detailed overview per team per subtask is available in Figure 3 in supplementary document B.

4.2. Model design predictors

Of all the teams that participated in both tasks, the challenge winner, SMU-MedVision, was the only team to implement two different model architectures for each task. Most teams used the same model architecture for the brain and the pelvis but trained it separately for both regions. Therefore, the limited number of teams that chose similar or different models/parameters for the brain and pelvis did not allow for visible trends in the rankings (Tables 3 and 4). Nevertheless, teams that used one model conditioned on the anatomy consistently secured relatively high ranks for both tasks. Still, the team that trained one collective model without conditioning on the anatomical region ranked last in both tasks.

In addition, plain CNN decoder-encoder and GAN-based models were prevalent among the teams. However, the teams that placed first and third in task 1 and second and fourth in task 2 used transformer-based approaches. These transformers showed significantly better performance in both regions for task 1 and in the brain for task 2, achieving average SSIM values of 0.88 ± 0.03 for task 1 and 0.90 ± 0.03 for task 2 (Fig. 3). Following the transformers, CNN encoder–decoder models were the next best-performing, yielding SSIM values of 0.85 ± 0.04 and 0.89 ± 0.04 for tasks 1 and 2, respectively. Conversely, teams using GANs tended to rank lower (Tables 3 and 4), with SSIM values of 0.83 ± 0.07 for task 1 and 0.87 ± 0.05 for task 2. Notably, GANs showed a significant performance drop, especially for the pelvis cases (Fig. 3).

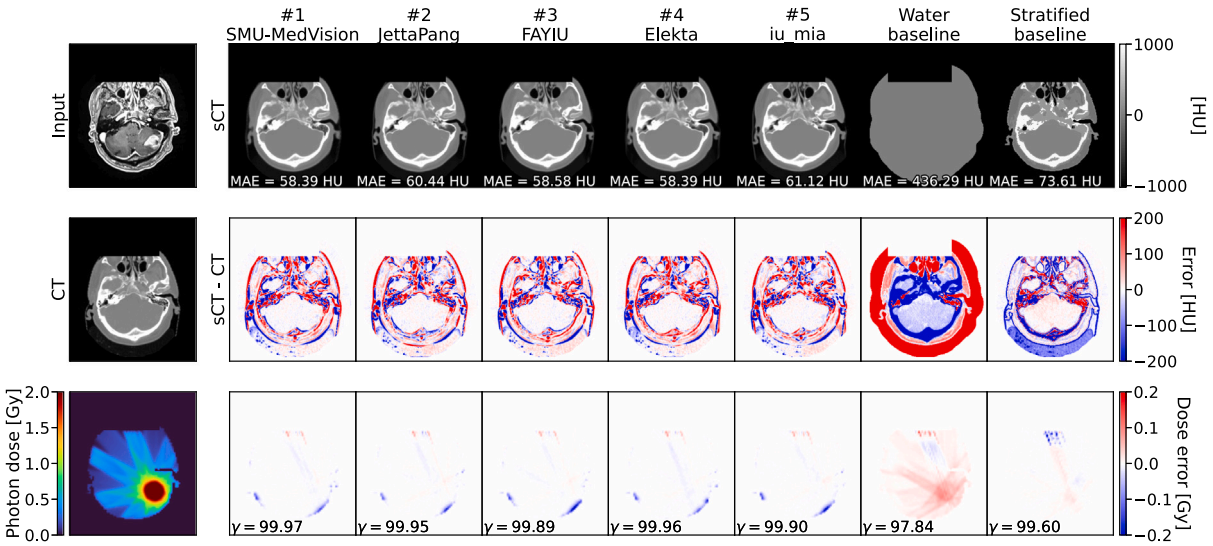
Table 5
All quantitative metrics ($\mu \pm \sigma$) produced by every participant in task 1 (MRI-to-CT) and task 2 (CBCT-to-CT). There were three image-based and six dose-based metrics: three for photon treatment and three for proton treatment. The best results per task per metric are marked in boldface.

Metric table 1: The quantitative metrics for task 1 (MRI-to-CT). Participants who scored worse than the water baseline on one image metric were excluded from the final ranking.

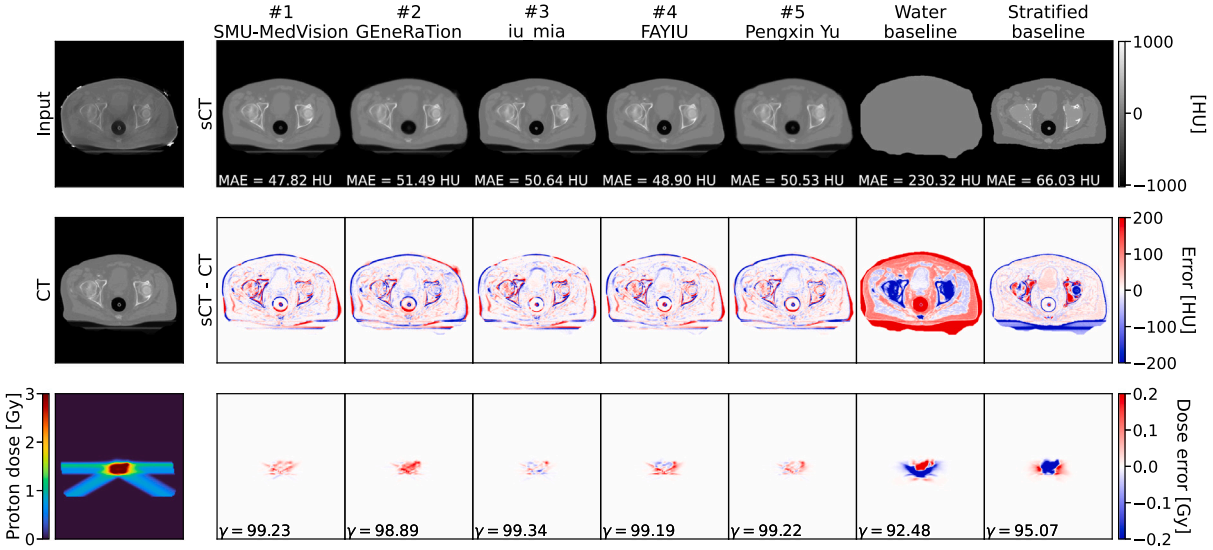
Rank	Team	Image metrics			Dose metrics					
					Photon			Proton		
		MAE (HU, ↓)	PSNR (dB, ↑)	SSIM (↑)	MAE (Gy, ↓)	DVH (↓)	$\gamma_{2\%/2\ mm}$ (↑)	MAE (Gy, ↓)	DVH (↓)	$\gamma_{2\%/2\ mm}$ (↑)
1	SMU-MedVision	**58.83 ± 13.41**	**29.61 ± 1.79**	**0.885 ± 0.029**	**0.0040 ± 0.0032**	**0.0265 ± 0.0382**	98.23 ± 4.45	0.0326 ± 0.0220	**0.2087 ± 0.2604**	97.28 ± 2.58
2	Jetta_Pang	65.73 ± 13.75	28.38 ± 1.68	0.869 ± 0.032	0.0040 ± 0.0037	0.0268 ± 0.0429	**99.00 ± 1.98**	**0.0316 ± 0.0194**	0.2208 ± 0.2680	**97.54 ± 2.37**
3	FAYIU	61.72 ± 13.32	28.83 ± 1.61	0.876 ± 0.030	0.0041 ± 0.0036	0.0273 ± 0.0437	98.18 ± 4.24	0.0320 ± 0.0202	0.2150 ± 0.2732	97.25 ± 2.56
4	Elekta	62.76 ± 13.06	28.80 ± 1.60	0.875 ± 0.030	0.0040 ± 0.0036	0.0286 ± 0.0525	98.15 ± 4.21	0.0332 ± 0.0220	0.2271 ± 0.2756	97.27 ± 2.50
5	iu_mia	62.83 ± 13.77	28.70 ± 1.59	0.873 ± 0.029	0.0040 ± 0.0034	0.0278 ± 0.0533	98.07 ± 4.65	0.0322 ± 0.0187	0.2173 ± 0.2608	97.31 ± 2.50
6	ShantouBME	67.54 ± 14.17	28.05 ± 1.55	0.863 ± 0.031	0.0042 ± 0.0034	0.0275 ± 0.0408	98.11 ± 4.18	0.0334 ± 0.0197	0.2185 ± 0.2575	97.23 ± 2.34
7	FGH_365	66.75 ± 13.18	28.59 ± 1.51	0.866 ± 0.032	0.0043 ± 0.0039	0.0324 ± 0.0598	97.94 ± 4.75	0.0349 ± 0.0254	0.2249 ± 0.2674	96.91 ± 2.82
8	USC-LONI	70.70 ± 14.40	27.77 ± 1.74	0.854 ± 0.034	0.0045 ± 0.0037	0.0331 ± 0.0590	97.97 ± 4.32	0.0365 ± 0.0210	0.2715 ± 0.3116	97.00 ± 2.77
9	UKA	77.39 ± 22.04	27.83 ± 2.16	0.849 ± 0.050	0.0045 ± 0.0036	0.0327 ± 0.0504	98.40 ± 4.04	0.0394 ± 0.0337	0.3641 ± 0.3683	96.79 ± 3.05
10	PSICPT_AI4PT	78.00 ± 27.44	27.55 ± 2.16	0.839 ± 0.049	0.0044 ± 0.0031	0.0316 ± 0.0487	98.17 ± 4.13	0.0370 ± 0.0263	0.2326 ± 0.2695	97.18 ± 2.10
11	SubtleCT	66.53 ± 14.63	28.51 ± 1.68	0.869 ± 0.029	0.0054 ± 0.0042	0.0370 ± 0.0624	98.20 ± 4.12	0.0443 ± 0.0308	0.2353 ± 0.2807	96.17 ± 2.99
12	mriG	82.01 ± 17.77	26.38 ± 1.53	0.842 ± 0.035	0.0046 ± 0.0040	0.0305 ± 0.0453	97.77 ± 4.24	0.0318 ± 0.0182	0.2363 ± 0.2772	97.08 ± 2.70
13	KoalAI	68.94 ± 11.82	28.14 ± 1.37	0.862 ± 0.029	0.0054 ± 0.0045	0.0338 ± 0.0432	97.94 ± 4.49	0.0460 ± 0.0314	0.2450 ± 0.2952	95.86 ± 3.60
14	Breizh-CT	93.57 ± 19.01	25.86 ± 1.43	0.806 ± 0.032	0.0050 ± 0.0041	0.0334 ± 0.0491	98.56 ± 2.44	0.0454 ± 0.0222	0.3052 ± 0.3153	95.97 ± 3.00
15	SKJP	88.42 ± 26.89	26.44 ± 2.03	0.815 ± 0.043	0.0063 ± 0.0047	0.0428 ± 0.0553	97.72 ± 4.97	0.0562 ± 0.0319	0.3445 ± 0.4111	94.83 ± 4.26
16	reza.karimzadeh	113.38 ± 20.35	24.71 ± 1.43	0.764 ± 0.034	0.0083 ± 0.0067	0.0542 ± 0.0644	97.02 ± 4.70	0.0565 ± 0.0288	0.4068 ± 0.6937	94.47 ± 3.80
17	thomashelfer	126.32 ± 17.01	23.69 ± 0.94	0.756 ± 0.029	0.0098 ± 0.0070	0.0599 ± 0.0643	97.38 ± 4.86	0.0791 ± 0.0585	0.3823 ± 0.4215	94.53 ± 3.66
18	X-MAN	117.88 ± 45.08	25.64 ± 2.20	0.774 ± 0.097	0.0117 ± 0.0148	0.0736 ± 0.0957	96.42 ± 5.00	0.0817 ± 0.0759	3514.188 ± 38491.0637	93.72 ± 5.77
19	Water baseline	332.93 ± 89.53	17.95 ± 1.73	0.552 ± 0.127	0.0166 ± 0.0104	0.0972 ± 0.1057	96.70 ± 5.36	0.1334 ± 0.0573	1.8218 ± 8.2652	88.87 ± 7.83
-	Stratified baseline	69.45 ± 16.33	28.42 ± 1.92	0.854 ± 0.027	0.0049 ± 0.0039	0.0347 ± 0.0504	98.21 ± 5.17	0.0515 ± 0.0309	0.4180 ± 0.3567	95.43 ± 3.76

Metric table 2: The quantitative metrics for task 2 (CBCT-to-CT). Participants who scored worse than the water baseline on one image metric were excluded from the final ranking.

Rank	Team	Image metrics			Dose metrics					
					Photon			Proton		
		MAE (HU, ↓)	PSNR (dB, ↑)	SSIM (↑)	MAE (Gy, ↓)	DVH (↓)	$\gamma_{2\%/2\ mm}$ (↑)	MAE (Gy, ↓)	DVH (↓)	$\gamma_{2\%/2\ mm}$ (↑)
1	SMU-MedVision	**49.95 ± 11.78**	**30.79 ± 2.00**	**0.906 ± 0.036**	**0.0038 ± 0.0042**	**0.0240 ± 0.0703**	99.49 ± 1.65	**0.0283 ± 0.0251**	**0.1663 ± 0.2235**	**97.57 ± 3.12**
2	GEneRaTion	55.50 ± 11.00	30.48 ± 1.72	0.897 ± 0.033	0.0040 ± 0.0042	**0.0241 ± 0.0530**	99.55 ± 1.20	0.0294 ± 0.0251	0.1689 ± 0.2188	97.42 ± 3.11
3	iu_mia	50.79 ± 11.81	30.58 ± 1.95	**0.906 ± 0.034**	0.0045 ± 0.0083	0.0326 ± 0.1543	98.99 ± 4.57	0.0336 ± 0.0408	0.1728 ± 0.2269	97.00 ± 4.72
4	FAYIU	51.18 ± 11.34	30.40 ± 1.93	0.903 ± 0.034	0.0044 ± 0.0080	0.0317 ± 0.1415	99.06 ± 4.29	0.0333 ± 0.0393	0.1711 ± 0.2197	97.09 ± 4.42
5	Pengxin Yu	54.05 ± 12.30	30.56 ± 1.95	0.900 ± 0.037	0.0043 ± 0.0070	0.0304 ± 0.1297	99.19 ± 3.72	0.0320 ± 0.0342	0.1762 ± 0.2296	97.16 ± 4.18
6	FGZ Medical Research	60.65 ± 12.56	29.67 ± 1.71	0.879 ± 0.039	0.0040 ± 0.0032	0.0251 ± 0.0442	**99.57 ± 1.07**	0.0307 ± 0.0208	0.2239 ± 0.2718	97.46 ± 2.85
7	KoalAI	56.13 ± 12.06	30.11 ± 1.89	0.897 ± 0.034	0.0055 ± 0.0080	0.0385 ± 0.1411	98.99 ± 4.38	0.0408 ± 0.0373	0.2106 ± 0.2497	96.05 ± 4.77
8	FGH_365	56.29 ± 11.08	30.24 ± 1.79	0.896 ± 0.035	0.0058 ± 0.0084	0.0410 ± 0.1517	98.97 ± 4.51	0.0432 ± 0.0439	0.2029 ± 0.2387	95.94 ± 5.08
9	UKA	65.46 ± 19.25	29.13 ± 2.64	0.881 ± 0.041	0.0049 ± 0.0092	0.0364 ± 0.1540	98.98 ± 4.89	0.0365 ± 0.0390	0.218 ± 0.2479	96.83 ± 4.79
10	Breizh-CT	71.28 ± 13.60	28.43 ± 1.65	0.863 ± 0.041	0.0052 ± 0.0052	0.0347 ± 0.0739	99.25 ± 2.23	0.0449 ± 0.0380	0.2309 ± 0.2495	96.06 ± 4.15
11	MedicalMind	68.40 ± 13.48	29.18 ± 1.63	0.875 ± 0.030	0.0094 ± 0.0110	0.0665 ± 0.1671	98.42 ± 5.38	0.0678 ± 0.0569	0.2395 ± 0.2534	94.11 ± 6.23
12	RRRocket_Lollies	71.58 ± 13.79	28.34 ± 1.50	0.862 ± 0.036	0.0099 ± 0.0095	0.0651 ± 0.1497	98.42 ± 4.95	0.0740 ± 0.0481	0.2744 ± 0.2678	92.32 ± 5.87
13	SKJP	78.63 ± 18.88	27.98 ± 1.71	0.853 ± 0.033	0.0113 ± 0.0092	0.0784 ± 0.1478	98.67 ± 4.76	0.0844 ± 0.0533	0.4572 ± 0.4907	91.50 ± 6.38
14	X-MAN	99.15 ± 59.43	27.51 ± 3.41	0.831 ± 0.087	0.0227 ± 0.0384	0.1150 ± 0.2003	92.44 ± 14.5	0.1035 ± 0.1056	0.3747 ± 0.3803	91.70 ± 10.8
15	Water baseline	344.26 ± 125.32	17.97 ± 2.08	0.546 ± 0.149	0.0191 ± 0.0118	0.1255 ± 0.1663	96.33 ± 5.80	0.1453 ± 0.0525	28.7704 ± 309.6595	85.08 ± 9.78
-	Stratified baseline	69.99 ± 18.93	28.65 ± 2.25	0.837 ± 0.057	0.0046 ± 0.0027	0.0332 ± 0.0432	**99.86 ± 0.46**	0.0432 ± 0.0282	0.3936 ± 0.3608	95.93 ± 2.97



(a) Example sCTs of participants for task 1 (MRI-to-CT). Despite significant synthesis errors at the tissue boundaries, high gamma-pass rates are achieved for all methods.



(b) Example sCTs of participants for task 2 (CBCT-to-CT). The largest synthesis errors occur at the boundaries between tissue types (e.g., air/body contour boundary or soft-tissue/bone boundary). Moreover, the anatomy does not fit in the FOV of the CBCT, requiring synthesis of tissue outside the FOV

Fig. 2. Examples of synthetic CTs for task 1 (MRI-to-CT; a) and task 2 (CBCT-to-CT: b). The model input is shown in the upper left, and the ground truth is in the center-left. The sCT of the top five participants for task 1 and task 2 are shown in the top row. The difference from ground truth CT is shown in the middle row. On the bottom left is the planned irradiation based on the CT for a photon (a) and proton (b) plan. The bottom row shows the dose difference when the treatment plan is applied to the sCT (CT dose - sCT dose). All values outside the body contour were masked.

Finally, the diffusion model, rarely adopted in this challenge, achieved SSIM values of 0.82 ± 0.06 and 0.88 ± 0.04 for tasks 1 and 2, respectively.

Only one team (RRRocket_Lollies, task 2) implemented an unsupervised approach which placed them close to the bottom of the ranking (12th out of 14). Due to the lack of unsupervised methods we could not extend the analysis on the supervision level.

Spatial configuration exhibited opposing trends in the two tasks (Fig. 4). For task 1 (MRI-to-CT), 3D patch-based, 2.5D, and 2D models achieved SSIM values of 0.83 ± 0.06, 0.85 ± 0.04, and 0.87 ± 0.03, respectively, with 2D models significantly outperforming the others. In contrast, for task 2 (CBCT-to-CT), 3D patch-based models (significantly) outperformed other models, with SSIM values of 0.89 ± 0.06, 0.88 ± 0.04, and 0.88 ± 0.04 for 3D patch-based, 2.5D, and 2D models, respectively.

No significant differences in choices for preprocessing, data augmentation, and postprocessing and, consequently, no trends in ranking were observed (Tables 3 and 4). The numerous combinations of processing steps and substantial differences in model design prevent definitive conclusions about the importance of specific processing steps.

4.3. Data influence

The image quality of the brain patients was significantly different between the centers (Fig. 5). In task 1, the participants generated sCTs for centers A, B, and C with an SSIM of 0.857 ± 0.052, 0.831 ± 0.056, and 0.852 ± 0.050, respectively. For task 2, the participants generated sCTs for centers A, B, and C with an SSIM of 0.883 ± 0.039, 0.921 ± 0.034, and 0.897 ± 0.035, respectively. No statistically significant differences in image similarity were observed between centers for the pelvis data. The sCTs in task 2 showed a better image similarity than those in task 1, with an MAE of 79.40 ± 28.30 HU for task 1 versus 63.50 ± 24.34 HU for task 2. When considering the dose metrics for brain cases in task 1, center B ($\gamma_{\text{photon}} = 92.03 \pm 6.84$) underperforms compared to centers A ($\gamma_{\text{photon}} = 99.65 \pm 1.09$) and C ($\gamma_{\text{photon}} = 99.93 \pm 0.17$). On the other hand, for pelvis cases in task 1, center A ($\gamma_{\text{photon}} = 98.29 \pm 3.00$) underperforms relative to center C ($\gamma_{\text{photon}} = 99.55 \pm 0.58$). For brain cases in task 2, minor dose differences were observed between the centers, with

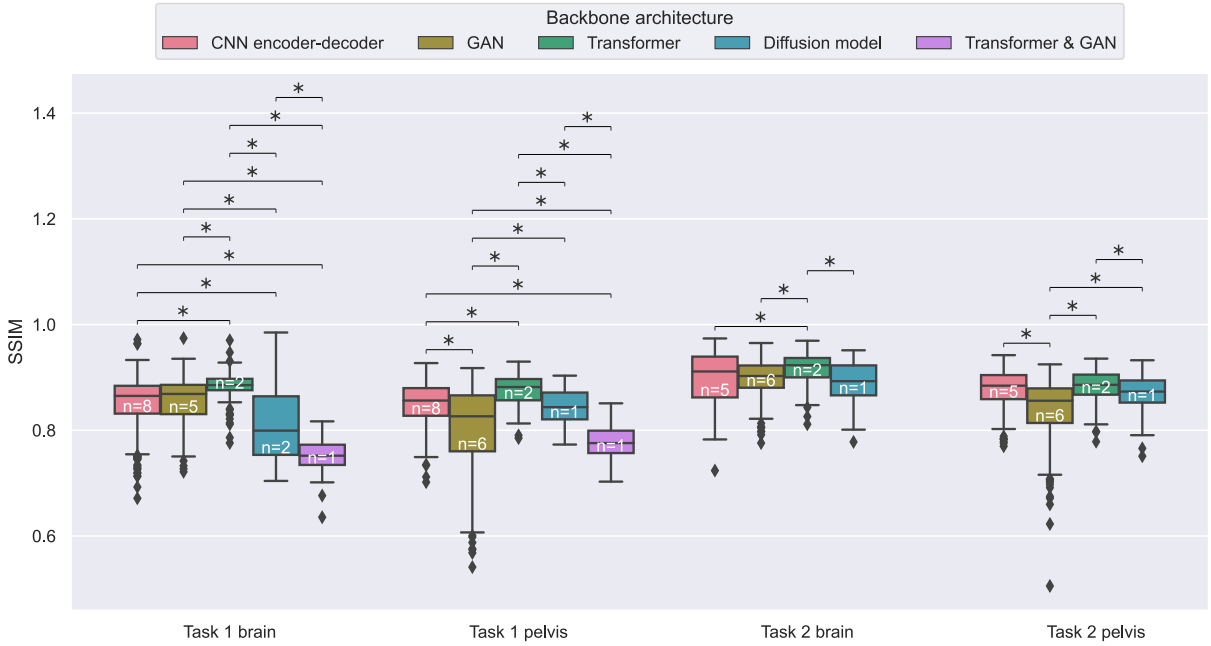


Fig. 3. Boxplots of SSIM values for all patients in each subtask, i.e., task 1 (MRI-to-CT) or 2 (CBCT-to-CT) and brain or pelvis, grouped by the model backbone choice of each team. The n in the boxes indicates the number of teams represented in that box. An asterisk indicates significant differences within one subtask.

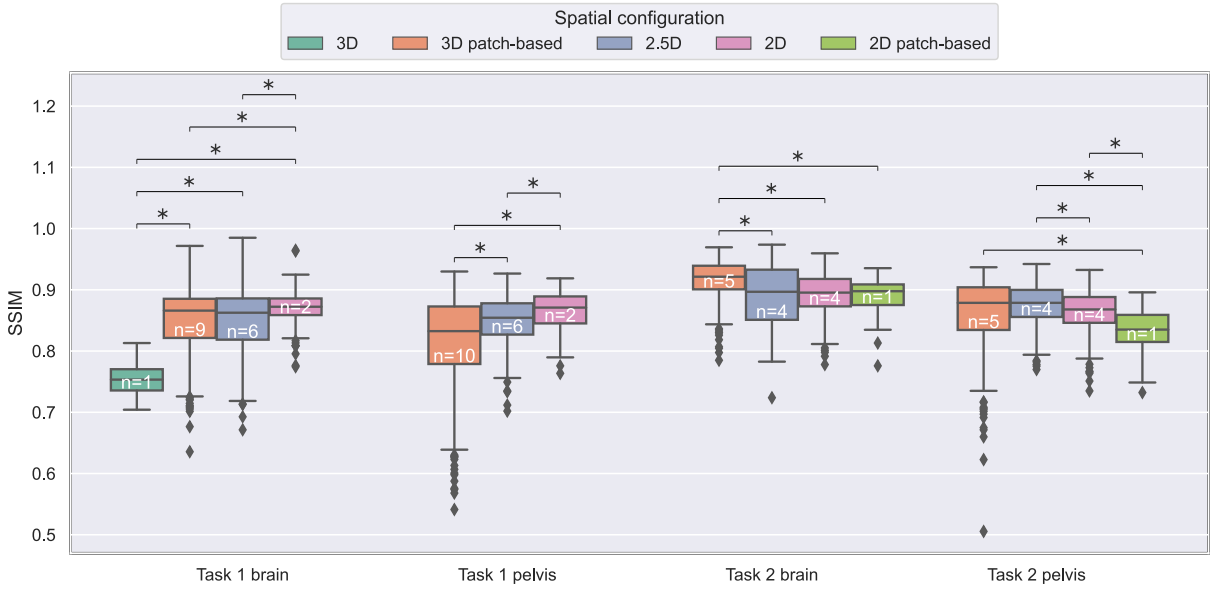


Fig. 4. Boxplots of SSIM values for all patients in each subtask, i.e., task 1 (MRI-to-CT) or 2 (CBCT-to-CT) and brain or pelvis, grouped by the spatial configuration of the models designed by the team. The n in the boxes indicates the number of teams represented in that box. An asterisk indicates significant differences within one subtask.

γ_{proton} = 98.80 ± 2.83, 96.87 ± 4.75, and 97.34 ± 4.94 for centers A, B, and C, respectively.

For task 1, each center employed consistent MRI scanning protocols for each anatomical region. Consequently, a comparison at the level of the MRI scan sequence yields identical results, as illustrated in Fig. 5. Moreover, the absence of variability in magnetic field strengths for centers B and C constrained this analysis to center A (Figure 4 in supplementary document B). For the brain, the only significant difference was observed for γ_{photon}, which decreased from 98.99 ± 1.43 for 1.5T to 97.33 ± 3.23 for 3T. In contrast, for the pelvis, a significant increase in performance was observed for 3T compared to 1.5T. Specifically, the SSIM increased from 0.83 ± 0.05 to 0.84 ± 0.05, γ_{photon} increased from 97.51 ± 3.45 to 98.75 ± 2.59, and γ_{proton} increased from 93.29 ± 4.05 to 95.64 ± 3.42.

A further investigation of the performance at the patient level, including a visual analysis of outlier patients, is presented in section 1.2 of supplementary document B.

4.4. Metric correlations

Fig. 6 highlights the correlations within the three metric groups. We observe strong correlations within the image similarity metric group, with the absolute inter-metric Spearman correlation coefficients $|\rho|$ ranging from 0.88 to 0.96 (Fig. 7). These values consistently measure the underlying aspects of all three image metrics. In contrast, the photon and proton metrics show weaker correlations within their groups. Among the dose metrics, the MAE_{dose} (photon) shows the highest correlation with the other dose metrics, such as γ pass rate, with

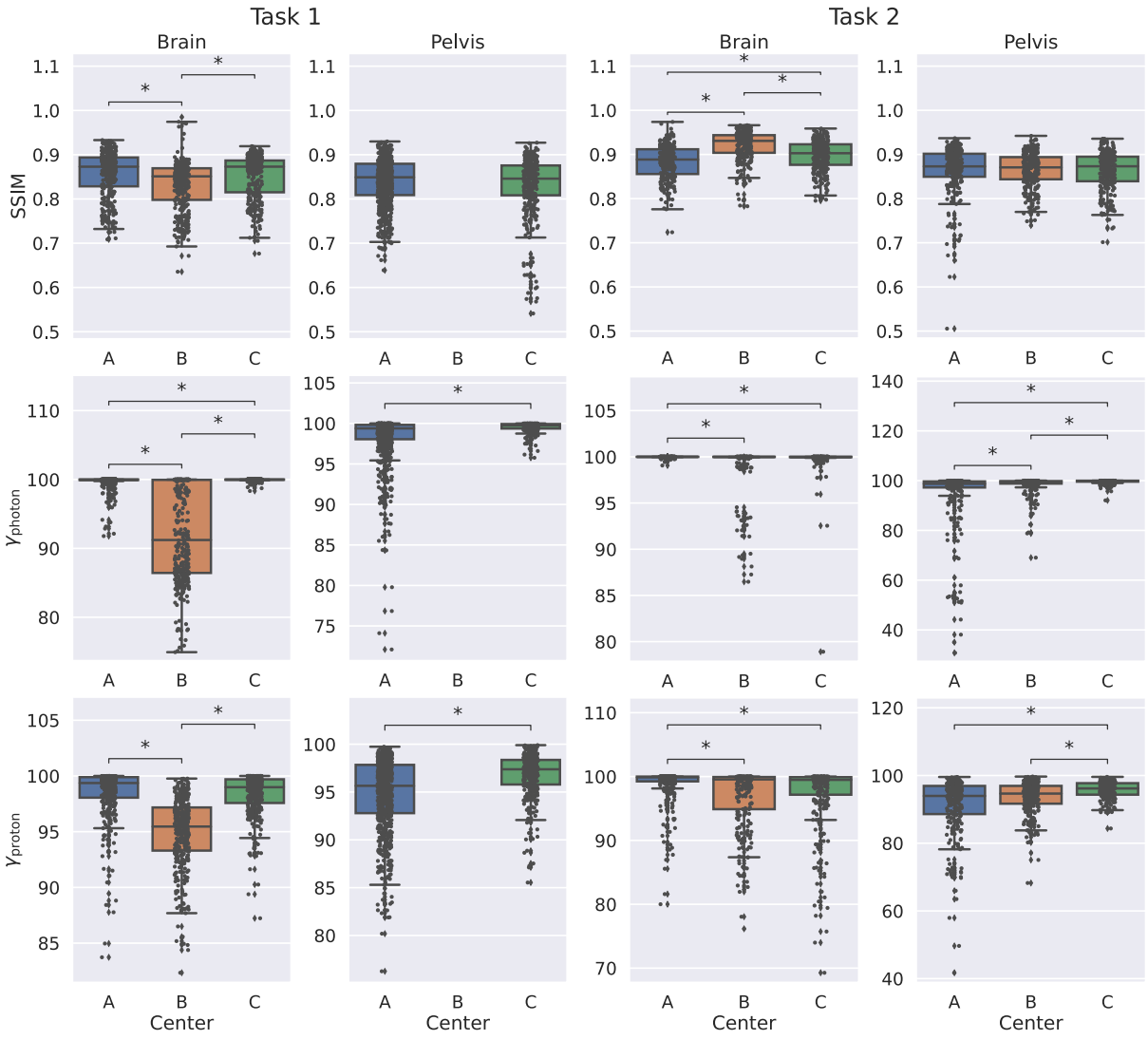
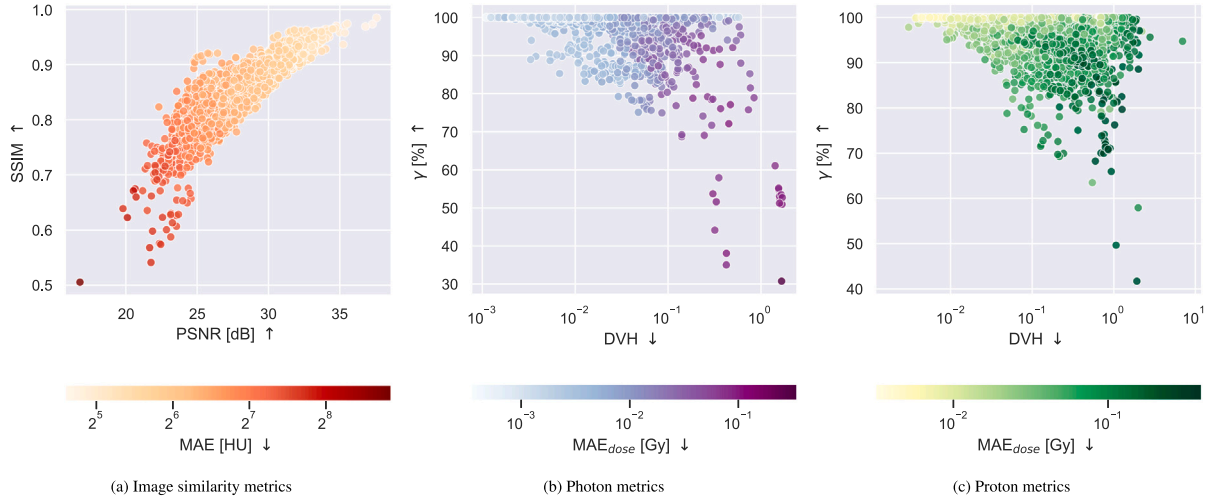


Fig. 5. Boxplots of the teams' performance in terms of SSIM and gamma pass rates for photon and proton, grouped by different subsets in our dataset, analyzing the differences between task, anatomical region and center. Asterisks indicate significant differences.



(a) Image similarity metrics (b) Photon metrics (c) Proton metrics

Fig. 6. Correlation plots among metrics in the three categories: (a) image metrics, (b) photon metrics, and (c) proton metrics. Each data point indicates a team's performance for one patient in either task 1 or 2. Note that some metrics are presented using a logarithmic scale, and one extreme outlier for the proton DVH metric (in the order of 1×10^5) is excluded from the plot.

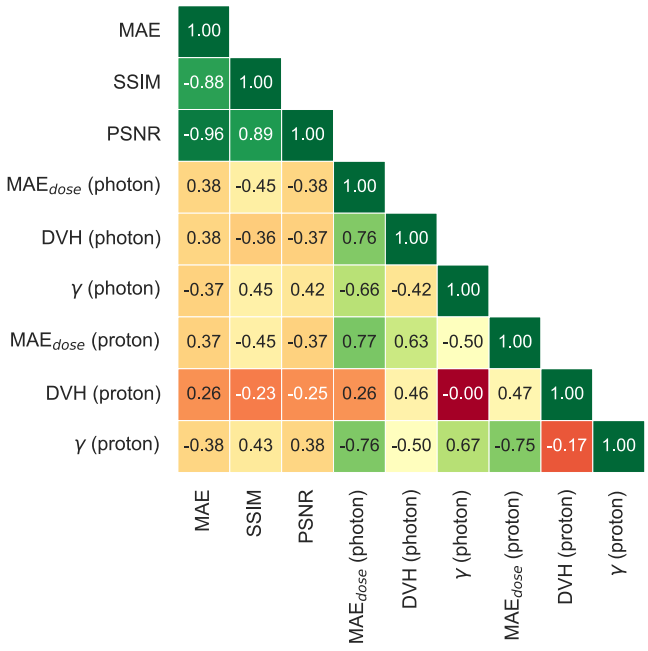


Fig. 7. Spearman rank correlation coefficient ρ between the different metrics. Note that the interpretation of the correlation coefficient is contingent upon whether both compared metrics exhibit concordant trends.

coefficients of −0.66 and −0.75 for photons and protons, respectively. While the correlation between the MAE_{dose} (photon) and DVH (photon) is strong (0.76), the correlation between MAE_{dose} (photon) and DVH (proton) is significantly lower (0.26). The proton DVH metric shows poor correlation with all metrics, highlighting the complex relationship between these metrics (Figs. 6(b), 6(c) and 7).

Furthermore, the metric groups correlate moderately with each other. The average absolute coefficients between image similarity metrics and photon metrics were 0.40 ± 0.03, while those with proton metrics were 0.47 ± 0.08. Moreover, the average absolute correlation coefficient between photon and proton metrics was 0.50 ± 0.23, with the large standard deviation introduced by a correlation coefficient of zero between DVH (proton) and γ (photon). Overall, the results strongly suggest that an sCT similar to the ground truth CT does not directly translate into a dose distribution similar to the reference distribution, highlighting that the different metrics focus on different aspects in evaluating sCTs.

4.5. Ranking stability and correlations

Fig. 8 illustrates that the challenge winner also secured the top position for both tasks under the three other ranking approaches, and teams at the bottom of the rankings are also stable across the ranking approaches. However, the middle-ranked teams experience notable shifts. For task 1, transitioning from MeanThenRank to MedianThenRank caused substantial changes for UKA (9 → 13) and mriG (12 → 8). Conversely, task 2's largest shifts occurred when changing from MeanThenRank to RankThenMedian for FGZ Medical research (6 → 2) and iu_mia (3 → 6). Despite these variations, all approaches strongly correlated with MeanThenRank approach, as indicated by Kendall's τ correlation coefficient (Table 6).

Fig. 9 demonstrates that the final rankings (determined by MeanThenRank) were relatively stable. There was high confidence in top-performing teams securing higher ranks and underperforming teams obtaining lower ranks, with the teams showing a maximum shift of 4 and 3 positions for tasks 1 and 2, respectively. SMU-MedVision had a 63.7% certainty of being the winner for task 1 and 99.7% for task 2,

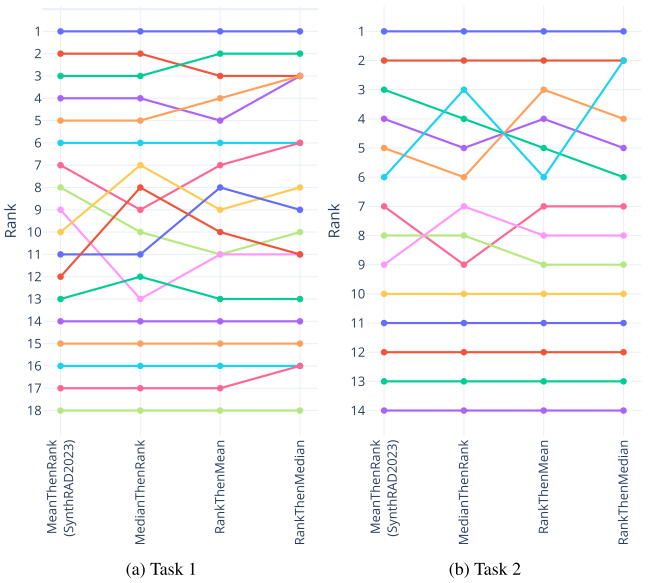


Fig. 8. Stability of the chosen ranking approach (MeanThenRank) compared to the other three. For all approaches, we used the mean over all test patients to obtain one average value per metric per team.

Table 6
Kendall's τ correlation coefficients for the ranking obtained from MeanThenRank compared to the other three ranking approaches.

Comparison	Task 1	Task 2
MeanThenRank vs. MedianThenRank	0.88	0.87
MeanThenRank vs. RankThenMean	0.88	0.91
MeanThenRank vs. RankThenMedian	0.91	0.84

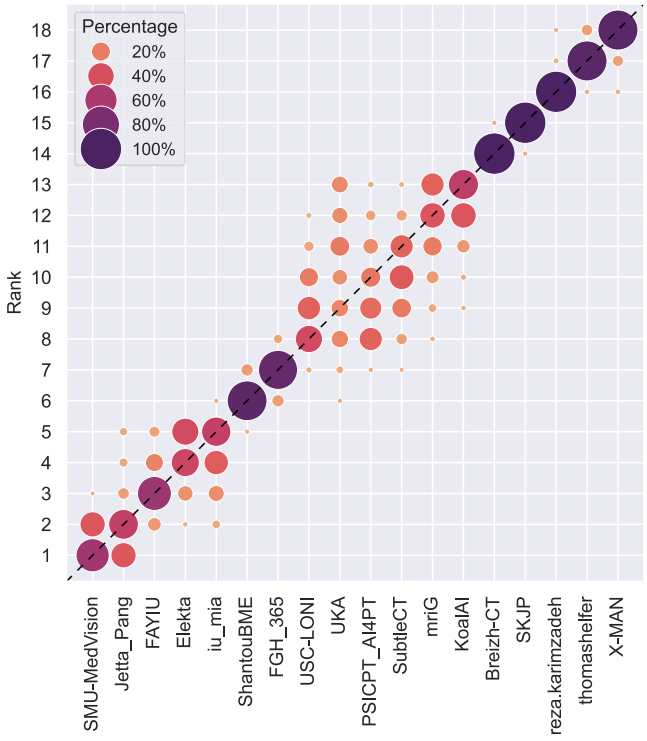
while certainties for the second to fifth places were lower, ranging from 45.0% to 66.7% for task 1 and from 30.5% to 60.4% for task 2. Teams in the middle of the rankings again showed some level of uncertainty, while it was inevitable that teams at the bottom of the ranking received the correct rank. This specifically holds for the last five teams for task 1 and the last six teams for task 2, with average certainties of $97.0 \pm 3.5\%$ and $99.8 \pm 0.4\%$, respectively.

5. Discussion

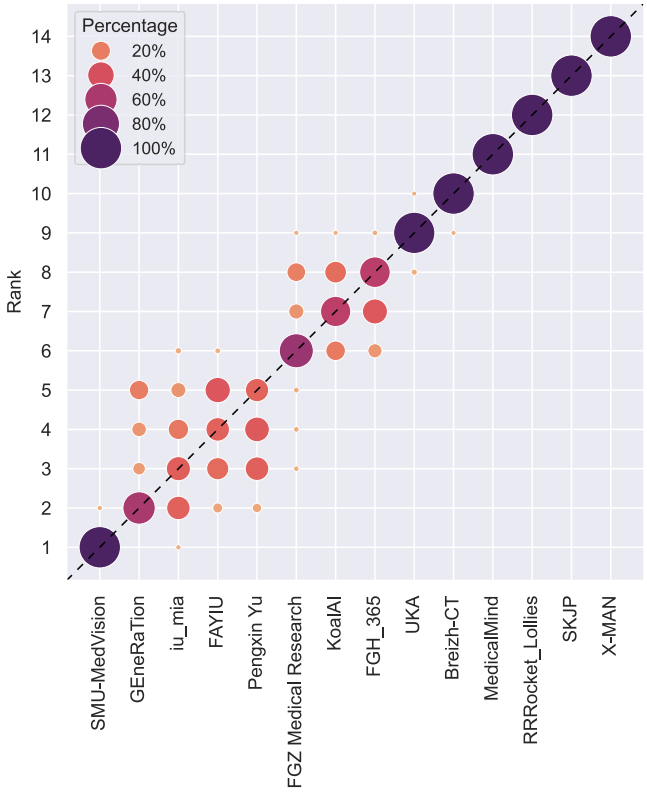
SynthRAD2023 allowed the comparison of deep learning techniques for synthesizing CT from MRI or CBCT. It is the first large-scale, multi-center challenge for generating *in-vivo* synthetic CT and garnered significant participation among the community, consisting of 617 participants, generating 39 valid submissions. The participants were generally able to synthesize high-quality sCT, outperforming the baseline algorithms in terms of image quality and dose accuracy.

The top five teams performed well, with SSIM values of at least 0.87 and 0.90 for tasks 1 and 2, respectively. Additionally, they exceeded gamma pass rates (2 mm/2%) of 98.07% for photon and 97.25% for proton treatment plans in task 1, and at least 98.99% for photon and 97.00% for proton treatment plans in task 2. These results indicate a high level of correspondence to the ground truth CTs. Nevertheless, despite the excellent performance, challenges remain for image synthesis. Difficulties for MRI-to-CT synthesis were encountered at air-tissue boundaries, potentially due to low MRI signal and magnetic susceptibility artifacts (Krupa and Bekiesińska-Figatowska, 2015). Additionally, in our dataset, the limited field-of-view of CBCT compared to CT introduced challenges in accurately synthesizing the complete body contour in the sCT.

Our analysis revealed that transformers (Vaswani et al., 2017) outperform CNN encoder–decoder models (e.g., U-Net Ronneberger



(a) Task 1



(b) Task 2

Fig. 9. Visualization of ranking stability. Blob size is proportional to the frequency of the rank achieved based on bootstrapping (N = 1000).

et al., 2015), which in turn outperform GANs (Goodfellow et al., 2014). Notably, recent architectures like diffusion models (Ho et al., 2020) and transformer-GAN combinations performed worse than the architectures mentioned above. These findings contrast recent reviews considering sCT generation (Spadea et al., 2021; Dayarathna et al., 2023), which either found no correlation between model architecture and performance or suggested that diffusion models hold promise in this field. Despite the statistically significant performance differences observed in our challenge, the differences were marginal, and the sample size was limited. In addition to comparing model architectures, it is important to acknowledge the potential impact of variations in training methodologies, including the reliability of hyperparameter search, on the observed performance differences among different approaches. Therefore, whether the observed differences stem solely from architectural choices or are significantly influenced by other aspects of the complex end-to-end pipeline, including preprocessing, data augmentation, postprocessing, and training procedures, remains inconclusive. For instance, previous literature suggests that data augmentation generally benefits generative models, suggesting that this step may play an essential role in model performance (Taylor and Nitschke, 2018; Steiner et al., 2021).

The 2D models outperformed 2.5D and (patch-based) 3D models for MRI-to-CT synthesis, while the 3D models outperformed the 2(.5)D approaches for CBCT-to-CT synthesis. These results hold for both pelvis and brain cases in both tasks. However, it has been shown that for MRI-to-CT synthesis 2.5D (multi-view) models outperform 2D models (Spadea et al., 2019; Maspero et al., 2020), and that 3D models outperform 2D models (Sun et al., 2022). We did not identify the cause of these contrasting results between MRI-to-CT and CBCT-to-CT synthesis. Future work could investigate why the impact of spatial dimension differs between these imaging modalities for synthetic CT generation.

We found that the image similarity metrics are highly correlated among themselves ($|\rho| \geq 0.88$) and that the MAE of the photon and proton dose distribution are moderately correlated to their respective gamma pass rates ($|\rho| \geq 0.66$). Specifically, the photon and proton DVH metrics are weakly correlated with the respective gamma pass rates ($|\rho| \leq 0.42$) (Fig. 7). Furthermore, the average correlations between the image similarity metrics and dose metrics are low ($|\rho| \leq 0.47$) despite the similar goal of measuring correspondence between the sCT and ground truth CT. The difference in correlations observed within and between metric groups may be attributed to the distinct regions where each metric is measured: image similarity was assessed within the dilated body contour, while dose metrics were calculated within high-dose regions or specific organs. These findings suggest that image similarity metrics should not be solely relied upon to determine the clinical suitability of a model, as they are not a reliable surrogate for clinically relevant dose metrics. Previous literature aligns with the finding, corroborating the poor correlation between image similarity and dose accuracy (Kieselmann et al., 2018; Peng et al., 2020). This highlights the need to perform thorough dose evaluations when clinically testing sCT generation approaches.

Two teams, i.e., UKA and PSCICP_4AI4PT, scored unexpectedly low in the final test phase compared to the validation phase due to implementation errors or misinterpreting data details. After (re-)opening the post-challenge phases, the two teams submitted corrected versions of the algorithms. Based on image similarity metrics alone, UKA climbed seven positions (9 → 2) in the rankings for both tasks. Similarly, PSCICP_4AI4PT climbed six positions (10 → 4) for task 1. During the open test phase, the teams could not resubmit their algorithms, as they ran successfully on the platform. The low scores of the erroneous algorithms underscore the fairness of the adopted rules.

5.1. Clinical impact

Despite the similarity between the sCTs generated by the participants and the ground truth CTs, there remains a lack of consensus regarding the criteria determining the clinical acceptability of an sCT (Vandewinckele et al., 2020). In radiotherapy, treatment planning

is defined to meet specific dose prescriptions and constraints. In this sense, dose-related metrics may be considered clinically significant. Some works have investigated clinical acceptance criteria for synthetic CTs. For example, Olberg et al. (2019) considered photon gamma pass rates greater than 98% acceptable using the 2 mm/2% criterion. On the other hand, Korsholm et al. (2014) proposes that treatments with a DVH difference of <2% are clinically acceptable. However, these criteria were proposed for breast, head-and-neck, and thorax sCT generation; it is unclear whether these criteria translate to different anatomical regions. Before addressing the clinical impact of the challenge results, it is crucial to consider the quality of the treatment plans adopted for SynthRAD2023 evaluation. Treatment planning techniques may differ between institutes. The planning techniques chosen have been based on constraints adopted in clinical guidelines (Hall et al., 2021; Lambrecht et al., 2018), making the results of the challenge of clinical relevance. The Linac and proton systems used in the treatment planning were generic; however, studies have demonstrated their effectiveness by showing that gamma pass rates deviate by a maximum of 0.5% when compared to dose engines adopted in clinical systems, independent of the irradiation type (Wieser et al., 2017).

To indicate clinically acceptable sCT, we propose considering an average gamma pass rate (2 mm/2%) above 99% and 97.5% for photon and proton irradiation in regions receiving at ≥10% of the prescribed dose, respectively. For the SynthRAD2023 challenge, only one team (Jetta_Pang) met these criteria for the MRI-to-CT task. For the CBCT-to-CT task, one team (SMU-MedVision) met both criteria. In contrast, five other teams (GEneRaTion, FAYIU, Pengxin Yu, FGZ Medical Research, and Breizh-CT) met only the photon criterion. As previously mentioned, the evaluation was affected by differences in patient positioning between the imaging sessions. Still, when considering the results on a population level, we do not expect to observe any systematic dose differences unless the sCT generation method introduced geometrically consistent distortion (Adjeiwaah et al., 2019). The lack of systematic dose differences suggests that the solutions offered by the participants are promising and of high quality. Before implementing any proposed clinical solutions, evaluating them according to the clinical standards specific to each facility using the commissioned treatment planning system is advisable.

Currently, commercial solutions to generate sCT from MRI or CBCT are available (Köhler et al., 2015; van Stralen et al., 2019; Cronholm et al., 2020; Archambault et al., 2020). It would be interesting to compare the algorithms submitted to the SynthRAD2023 challenge with these commercial solutions. Some of these commercial solutions require a dedicated imaging protocol to generate accurate sCT data (Florkow et al., 2020; Bratova et al., 2019; Liu et al., 2023), making the comparison challenging. Exploring the necessity of specialized imaging protocols, or in simpler terms, assessing the ability of sCT algorithms to generalize across different input variations as in Nijskens et al. (2023), could be worthwhile.

5.2. Limitations of the SynthRAD2023 dataset and setup

A substantial multi-center dataset was gathered for the SynthRAD2023 challenge. However, the dataset can be further improved despite its size and diversity. For example, the dataset consists solely of patients treated at Dutch hospitals, which may limit the dataset's heterogeneity, possibly resulting in low performance for case outliers in the data distributions. Additionally, it is important to acknowledge that the included MRIs represent only a subset of the magnetic field strengths commonly used in clinical practice (1.5T and 3T). This limits the generalizability of the findings to the broader spectrum of clinical MRI applications, where other field strengths are routinely employed. While the inclusion of three centers represents a commendable starting point, extending the dataset with international data may improve the generalization capabilities of the submitted models and increase the clinical impact.

A model ideally should generalize across different centers without conditional fine-tuning, as current commercial solutions are not center-specific. While some participants incorporated center-based prediction and optimization using information shared in the training and validation sets, effective models should extend beyond the provided centers to make a clinical impact. Future challenges may consider whether circumventing such information may lead to designing more general approaches.

Furthermore, the SynthRAD2023 dataset contained rigidly registered image pairs, resulting in residual anatomical mismatch after registration, as mentioned above. Reducing the registration error, e.g., recurring to deformable registration, may improve performances (Florkow et al., 2020). However, it may also confound possible geometrical distortion to the input images the models may introduce, which is undesirable in a clinical scenario (Pappas et al., 2017). The impact of residual misregistration is corroborated by the paired nature of the dataset, and could be mitigated by performing unpaired synthetic CT generation. However, unpaired sCT generation prohibits a dosimetric evaluation of the generated sCT, and limits the use of established image similarity metric, such as the SSIM or PSNR.

An additional dataset limitation stems from the automated processing pipeline, where, for all brain patients from center B in task 1, the treatment table was included within the dilated body contour. At the same time, the table was successfully excluded from the dataset by the other centers, leading to inconsistent table representation in the dose evaluations. Moreover, for two out of sixty pelvis patients in task 2, the field-of-view of the CBCT was smaller than the body contour. Such patients can be considered outliers, and for future challenges, it would be beneficial to revise case selection and exclude them from the test set. The inclusion of the table in the mask and limited CBCT field-of-view had minor impact on the image similarity evaluation, which was computed within the provided mask, but could be more substantial for the dose evaluation due to beam attenuation. Note that the inconsistency was present for all teams, leaving the challenge ranking unbiased.

Another limitation arose from the absence of dose evaluation during the validation phase, hindering teams from optimizing their models for this radiotherapy-related metric. On the other hand, the lack of dose metrics during validation may have compelled participants to develop general methods that could function irrespective of the chosen planning strategy.

5.3. Future direction

SynthRAD2023 has set out to advance the state-of-the-art in MRI-to-CT and CBCT-to-CT generation. While the results are promising, these tasks have not yet been solved during this challenge. The dataset only included brain and pelvis patients. Other, maybe more challenging, anatomical regions could benefit from sCT generation, such as the thorax, head-and-neck, breast, or abdomen (Spadea et al., 2021). In addition, it would be of interest to examine the generalizability of the models by including test data from centers that were not present in the training data (Texier et al., 2023).

The positive reception to SynthRAD2023 has spurred the development of SynthRAD2025, which aims to expand the challenge beyond the Dutch national domain into more unexplored anatomical regions, such as the head-and-neck and abdomen.

Furthermore, addressing the limitations in data preparation and image registration discussed earlier will enhance the analysis of future challenges.

Lastly, we anticipate that the post-challenge phases will offer opportunities to validate and enhance the statistical robustness of the challenge's conclusions, enabling other researchers to compare their methods with the results of SynthRAD2023.

6. Conclusion

While synthetic CT generation has already become a clinical reality (Spadea et al., 2021), the SynthRAD2023 Grand Challenge represents a pivotal advancement in image synthesis for radiotherapy planning. The challenge marks the first multi-center challenge with a substantial dataset, serving as a catalyst for further innovation in radiotherapy. Participants showcased their ability to generate high-quality sCTs, demonstrating high image similarity and accurate dose distributions. These achievements highlight the potential of deep learning for enhancing sCT generation. However, it is important to recognize that solely relying on image similarity metrics may not adequately capture the clinical applicability of sCTs. Nonetheless, these significant strides hold promise for reducing reliance on conventional CT and improving efficiency in radiotherapy.

CRediT authorship contribution statement

Evi M.C. Huijben: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Maarten L. Terpstra:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Arthur Jr. Galapon:** Writing – review & editing, Visualization, Resources, Investigation, Formal analysis, Data curation, Conceptualization. **Suraj Pai:** Writing – review & editing, Visualization, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Adrian Thummerer:** Writing – review & editing, Software, Resources, Methodology, Data curation, Conceptualization. **Peter Koopmans:** Writing – review & editing, Methodology, Conceptualization. **Manya Afonso:** Writing – review & editing, Funding acquisition, Conceptualization. **Maureen van Eijnatten:** Writing – review & editing, Funding acquisition, Conceptualization. **Oliver Gurney-Champion:** Writing – review & editing, Conceptualization. **Zeli Chen:** Writing – review & editing, Software, Methodology. **Yiwen Zhang:** Writing – review & editing, Software, Methodology. **Kaiyi Zheng:** Writing – review & editing, Software, Methodology. **Chuanpu Li:** Writing – review & editing, Software, Methodology. **Haowen Pang:** Writing – review & editing, Software, Methodology. **Chuyang Ye:** Writing – review & editing, Software, Methodology. **Runqi Wang:** Writing – review & editing, Software, Methodology. **Tao Song:** Writing – review & editing, Software, Methodology. **Fuxin Fan:** Writing – review & editing, Software, Methodology. **Jingna Qiu:** Writing – review & editing, Software, Methodology. **Yixing Huang:** Writing – review & editing, Software, Methodology. **Juhyung Ha:** Writing – review & editing, Software, Methodology. **Jong Sung Park:** Writing – review & editing, Software, Methodology. **Alexandra Alain-Beaudoin:** Writing – review & editing, Software, Methodology. **Silvain Bériault:** Writing – review & editing, Software, Methodology. **Pengxin Yu:** Writing – review & editing, Software, Methodology. **Hongbin Guo:** Writing – review & editing, Supervision, Methodology. **Zhanyao Huang:** Writing – review & editing, Software, Methodology. **Gengwan Li:** Writing – review & editing, Software, Methodology. **Xueru Zhang:** Writing – review & editing, Software, Methodology. **Yubo Fan:** Writing – review & editing, Software, Methodology, Conceptualization. **Han Liu:** Writing – review & editing, Software, Methodology. **Bowen Xin:** Writing – review & editing, Software, Methodology, Conceptualization. **Aaron Nicolson:** Writing – review & editing, Software, Methodology. **Lujia Zhong:** Writing – review & editing, Software, Methodology. **Zhiwei Deng:** Writing – review & editing, Software, Methodology. **Gustav Müller-Franzes:** Writing – review & editing, Software, Methodology. **Firas Khader:** Writing – review & editing, Software, Methodology. **Xia Li:** Writing – review & editing, Software, Methodology. **Ye Zhang:** Writing – review & editing, Software, Methodology. **Cédric Hémon:** Writing – review & editing, Software, Methodology. **Valentin Boussot:** Writing – review & editing, Software, Methodology. **Zhihao Zhang:** Writing – review & editing, Software, Methodology. **Long Wang:** Writing – review & editing, Software, Methodology. **Lu Bai:** Writing – review & editing, Software, Methodology. **Shaobin Wang:** Writing – review & editing, Software, Methodology. **Derk Mus:** Writing – review & editing, Software, Methodology. **Bram Kooiman:** Writing – review & editing, Software, Methodology. **Chelsea A.H. Sargeant:** Writing – review & editing, Software, Methodology. **Edward G.A. Henderson:** Writing – review & editing, Software, Methodology. **Satoshi Kondo:** Writing – review & editing, Software, Methodology. **Satoshi Kasai:** Writing – review & editing, Software, Methodology. **Reza Karimzadeh:** Writing – review & editing, Software, Methodology. **Bulat Ibragimov:** Writing – review & editing, Software, Methodology, Conceptualization. **Thomas Helfer:** Writing – review & editing, Software, Methodology. **Jessica Dafflon:** Writing – review & editing, Software, Methodology. **Zijie Chen:** Writing – review & editing, Software, Methodology. **Enpei Wang:** Writing – review & editing, Software, Methodology. **Zoltan Perko:** Writing – review & editing, Validation, Software, Data curation, Conceptualization. **Matteo Maspero:** Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal Analysis, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Matteo Maspero reports financial support was provided by EWUU alliance. Manya Afonso, Maureen van Eijnatten reports financial support was provided by EWUU alliance.

Runqi Wang, Tao Song, Zhihao Zhang, Long Wang reports a relationship with Subtle Medical that includes: employment.

Alexandra Alain-Beaudoin, Silvain Bériault reports a relationship with Elekta Ltd that includes: employment. Alexandra Alain-Beaudoin, Silvain Beriault has patent pending to Elekta Ltd.

Pengxin Yu reports a relationship with Infervision Medical Technology Co., Ltd that includes: employment.

Gengwan Li, Xueru Zhang reports a relationship with Canon Medical Systems Corporation that includes: employment.

Lujia Zhong, Zhiwei Deng reports financial support was provided by University of Southern California.

Valentin Boussot reports financial support was provided by French National Research Agency. Valentin Boussot reports financial support was provided by Rennes Métropole. Cédric Hémon reports financial support was provided by Elekta AB. Cédric Hémon reports financial support was provided by CominLabs. Valentin Boussot reports a relationship with University of Rennes that includes: employment. Cédric Hémon reports a relationship with University of Rennes that includes: employment. Cédric Hémon reports a relationship with Centre Eugène Marquis that includes: employment.

Lu Bai, Shaobin Wang reports a relationship with MedMind Technology Co. Ltd that includes: employment. Lu Bai and Shaobing Wang had adopted some training ideas from Du Yi who is a staff in Institute of Medical Technology, Peking University Health Science Center, Beijing, China

Derk Mus, Bram Kooiman reports a relationship with MRI Guidance BV that includes: employment.

Chelsea Sargeant reports a relationship with Elekta AB that includes: funding grants and travel reimbursement. Edward Henderson reports a relationship with Manchester Cancer Research Centre that includes: funding grants and travel reimbursement.

Zijie Chen, Enpei Wang reports a relationship with Shenying Medical Technology (Shenzhen) Co., Ltd that includes: employment.

If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Training data, validation input, preprocessing, and evaluation code are publicly available. Ground truth validation and test data will be made available at the closure of the challenge (2028).

Declaration of generative AI

While preparing this work, the authors used Writefull, DeepL Write, and ChatGPT to enhance the writing structure and refine grammar. After using these tools, the authors reviewed and edited the content as needed and took full responsibility for the publication's content.

Acknowledgments

The organizers would like to thank Eric van der Bijl for leading the data collection at Radboud UMC Nijmegen, Cornelis (Nico) AT van den Berg for supporting the challenge offering facilities and hosting the administrative account within UMC Utrecht, Joost JC Verhoeff for supporting the data retrieval at UMC Utrecht, Stefan Both supporting the data retrieval at UMC Groningen. We thank the European Society of Radiation Oncology (ESTRO), https://www.estro.org/ and the Nederlandse Vereniging voor Klinische Fysica (NVKF) https://www.nvkf.nl/ for endorsing the event.

Funding

SynthRAD2023 was funded by the Seed Fund granted TU/e, WUR, UU, UMC Utrecht (EWUU) alliance (https://ewuu.nl) with the Seed Fund round of November 2022.

Appendix A. Participation rules and prize policies

To ensure fairness and transparency in SynthRAD2023, organizers, data providers, and contributors were prohibited from participating in the challenge since data providers and organizers had access to the data, including the test set ground truth CTs. However, members affiliated with the organizers' institutes were allowed to participate, provided they had not co-authored any publications with the organizers in the year preceding the challenge.

Participants were required to develop fully automated methods that run in the Amazon Web Services (AWS) cloud environment using a single `g4dn.2xlarge` instance. This instance includes a GPU with 16 GB VRAM, an 8-core CPU, and 32 GB RAM. In this environment, the inference time for generating an sCT for a single case (one patient) is constrained to a maximum of 15 min.

Teams receiving a prize had to present their methodology at MICCAI 2023, sign all necessary prize acceptance documents, and submit a detailed paper in LNCS format outlining their methods. Additionally, participants committed to citing both the data challenge paper (Thummerer et al., 2023a) and this challenge overview paper in subsequent publications, whether scientific or non-scientific. Although sharing codes was strongly encouraged, it was not mandatory. The challenge results and rankings were publicly announced after the test phase concluded. The top five teams for both tasks were awarded a total of €10,000, with the following distribution: €2200, €1250, €850, €500, €200.

The complete challenge design can be found at Thummerer et al. (2023b).

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.media.2024.103276.

References

Adjeiwaah, M., Bylund, M., Lundman, J.A., Söderström, K., Zackrisson, B., Jonsson, J.H., Garpebring, A., Nyholm, T., 2019. Dosimetric impact of MRI distortions: A study on head and neck cancers. Int. J. Radiat. Oncol. Biol. Phys. 103 (4), 994–1003. http://dx.doi.org/10.1016/j.ijrobp.2018.11.037.

Archambault, Y., Boylan, C., Bullock, D., Morgas, T., Peltola, J., Ruokokoski, E., Genghi, A., Haas, B., Suhonen, P., Thompson, S., 2020. Making on-line adaptive radiotherapy possible using artificial intelligence and machine learning for efficient daily re-planning. Med. Phys. Intl. J. 8 (2).

Bratova, I., Paluska, P., Grepl, J., Sykorova, P., Jansa, J., Hodek, M., Sirak, I., Vosmik, M., Petera, J., 2019. Validation of dose distribution computation on sCT images generated from MRI scans by Philips MRCAT. Rep. Pract. Oncol. Radiother. 24 (2), 245–250.

Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A., 2022. Monai: An open-source framework for deep learning in healthcare. arXiv preprint arXiv:2211.02701.

Chandra, R.A., Keane, F.K., Voncken, F.E.M., Thomas, C.R., 2021. Contemporary radiotherapy: present and future. Lancet 398 (10295), 171–184. http://dx.doi.org/10.1016/s0140-6736(21)00233-6.

Chatterjee, S., Sciarra, A., Dünnwald, M., Mushunuri, R.V., Podishetti, R., Rao, R.N., Gopinath, G.D., Oeltze-Jafra, S., Speck, O., Nürnberger, A., 2021. ShuffleUNet: Super resolution of diffusion-weighted MRIs using deep learning. In: 2021 29th European Signal Processing Conference. EUSIPCO, IEEE, pp. 940–944. http://dx.doi.org/10.23919/EUSIPCO54536.2021.9615963.

Chen, Z., Li, C., Zheng, K., Zhang, Y., Wu, Y., Feng, Q., Zhong, L., Yang, W., 2023. GLFA-NET: A hybrid network for Mr-To-Ct synthesis via global and local feature aggregation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging. ISBI, pp. 1–5. http://dx.doi.org/10.1109/ISBI53787.2023.10230486.

Cronholm, R.O., Karlsson, A., Siversson, C., 2020. MRI only radiotherapy planning using the transfer function estimation algorithm.

Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., Chen, Z., 2023. Deep learning based synthesis of MRI, CT and PET: Review and analysis. Med. Image Anal. 103046. http://dx.doi.org/10.1016/j.media.2023.103046.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. http://dx.doi.org/10.48550/ARXIV.2010.11929.

Edmund, J.M., Nyholm, T., 2017. A review of substitute CT generation for MRI-only radiation therapy. Radiat. Oncol. 12 (1), http://dx.doi.org/10.1186/s13014-016-0747-y.

Ezzell, G.A., Burmeister, J.W., Dogan, N., LoSasso, T.J., Mechalakos, J.G., Mihailidis, D., Molineu, A., Palta, J.R., Ramsey, C.R., Salter, B.J., Shi, J., Xia, P., Yue, N.J., Xiao, Y., 2009. IMRT commissioning: multiple institution planning and dosimetry comparisons, a report from AAPM task group 119. Med. Phys. 36 (11), 5359–5373. http://dx.doi.org/10.1118/1.3238104.

Florkow, M.C., Zijlstra, F., Willemsen, K., Maspero, M., van den Berg, C.A.T., Kerkmeijer, L.G., Castelein, R.M., Weinans, H., Viergever, M.A., van Stralen, M., Seevinck, P.R., 2020. Deep learning–based MR-to-CT synthesis: the influence of varying gradient echo–based MR images as input channels. Magn. Reson. Med. 83 (4), 1429–1441. http://dx.doi.org/10.1002/mrm.28008.

Ge, R., Yang, G., Xu, C., Chen, Y., Luo, L., Li, S., 2019. Stereo-correlation and noise-distribution aware ResVoxGAN for dense slices reconstruction and noise reduction in thick low-dose CT. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019. Springer, pp. 328–338. http://dx.doi.org/10.1007/978-3-030-32226-7_37.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative Adversarial Nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems. Vol. 27, Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

Grégoire, V., Mackie, T., 2011. State of the art on dose prescription, reporting and recording in intensity-modulated radiation therapy (ICRU report no. 83). Cancer/Radiothér. 15 (6–7), 555–559. http://dx.doi.org/10.1016/j.canrad.2011.04.003.

Hall, W.A., Paulson, E., Davis, B.J., Spratt, D.E., Morgan, T.M., Dearnaley, D., Tree, A.C., Efstathiou, J.A., Harisinghani, M., Jani, A.B., Buyyounouski, M.K., Pisansky, T.M., Tran, P.T., Karnes, R.J., Chen, R.C., Cury, F.L., Michalski, J.M., Rosenthal, S.A., Koontz, B.F., Wong, A.C., Nguyen, P.L., Hope, T.A., Feng, F., 2021. NRG oncology updated international consensus atlas on pelvic lymph node volumes for intact and postoperative prostate cancer. Int. J. Radiat. Oncol. Biol. Phys. 109 (1), 174–185. http://dx.doi.org/10.1016/j.ijrobp.2020.08.034.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H.R., Xu, D., 2021. Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images. In: International MICCAI Brainlesion Workshop. Springer, pp. 272–284. http://dx.doi.org/10.1007/978-3-031-08999-2_22.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, pp. 16000–16009.

Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. arXiv preprint arXiv:2006.11239.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 65–70, https://www.jstor.org/stable/4615733.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18 (2), 203–211.

Isola, P., Zhu, J.-Y., Zhou, T., Efros, A.A., 2017. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR, pp. 1125–1134.

Johnson, J., Alahi, A., Fei-Fei, L., 2016. Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part II 14. Springer, pp. 694–711.

Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30 (1–2), 81–93. http://dx.doi.org/10.1093/biomet/30.1-2.81.

Kieselmann, J.P., Kamerling, C.P., Burgos, N., Menten, M.J., Fuller, C.D., Nill, S., Cardoso, M.J., Oelfke, U., 2018. Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. Phys. Med. Biol. 63 (14), 145007. http://dx.doi.org/10.1088/1361-6560/aacb65.

Köhler, M., Vaara, T., van Grootel, M., Hoogeveen, R., Kemppainen, R., Renisch, S., 2015. MR-Only Simulation for Radiotherapy Planning – Whitepaper: Philips MRCAT for Prostate Dose Calculations Using Only MRI Data. White Paper, Philips.

Korsholm, M.E., Waring, L.W., Edmund, J.M., 2014. A criterion for the reliable use of MRI-only radiotherapy. Radiat. Oncol. 9 (1), http://dx.doi.org/10.1186/1748-717x-9-16.

Krupa, K., Bekiesińska-Figatowska, M., 2015. Artifacts in magnetic resonance imaging. Polish J. Radiol. 80, 93.

Lagendijk, J.J.W., Raaymakers, B.W., Van den Berg, C.A.T., Moerland, M.A., Philippens, M.E., van Vulpen, M., 2014. MR guidance in radiotherapy. Phys. Med. Biol. 59 (21), R349–R369. http://dx.doi.org/10.1088/0031-9155/59/21/r349.

Lambrecht, M., Eekers, D.B., Alapetite, C., Burnet, N.G., Calugaru, V., Coremans, I.E., Fossati, P., Høyer, M., Langendijk, J.A., Méndez Romero, A., Paulsen, F., Perpar, A., Renard, L., de Ruysscher, D., Timmermann, B., Vitek, P., Weber, D.C., van der Weide, H.L., Whitfield, G.A., Wiggenraad, R., Roelofs, E., Witt Nyström, P., Troost, E.G.C., 2018. Radiation dose constraints for organs at risk in neuro-oncology; the European particle therapy network consensus. Radiother. Oncol. 128 (1), 26–36. http://dx.doi.org/10.1016/j.radonc.2018.05.001.

Liang, J., Cao, J., Sun, G., Zhang, K., Van Gool, L., Timofte, R., 2021. SwinIR: Image restoration using swin transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1833–1844.

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., Wei, F., Guo, B., 2022. Swin transformer v2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10012–10022.

Liu, H., Schaal, D., Curry, H., Clark, R., Magliari, A., Kupelian, P., Khuntia, D., Beriwal, S., 2023. Review of cone beam computed tomography based online adaptive radiotherapy: Current trend and future direction. Radiat. Oncol. 18 (1), 144. http://dx.doi.org/10.1186/s13014-023-02340-2.

Low, D.A., Harms, W.B., Mutic, S., Purdy, J.A., 1998. A technique for the quantitative evaluation of dose distributions. Med. Phys. 25 (5), 656–661. http://dx.doi.org/10.1118/1.598248.

Mann, H.B., Whitney, D.R., 1947. On a test of whether one of two random variables is stochastically larger than the other. Ann. Math. Stat. 18 (1), 50–60. http://dx.doi.org/10.1214/aoms/1177730491.

Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S., 2017. Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802.

Maspero, M., Bentvelzen, L.G., Savenije, M.H., Guerreiro, F., Seravalli, E., Janssens, G.O., van den Berg, C.A., Philippens, M.E., 2020. Deep learning-based synthetic CT generation for paediatric brain MR-only photon and proton radiotherapy. Radiother. Oncol. 153, 197–204. http://dx.doi.org/10.1016/j.radonc.2020.09.029.

Maspero, M., Van den Berg, C.A., Landry, G., Belka, C., Parodi, K., Seevinck, P.R., Raaymakers, B.W., Kurz, C., 2017. Feasibility of MR-only proton dose calculations for prostate cancer radiotherapy using a commercial pseudo-CT generation method. Phys. Med. Biol. 62 (24), 9159. http://dx.doi.org/10.1088/1361-6560/aa9677.

Mitchell, G., 2013. The rationale for fractionation in radiotherapy. Clin. J. Oncol. Nurs. 17 (4), 412–417. http://dx.doi.org/10.1188/13.cjon.412-417.

Miyato, T., Kataoka, T., Koyama, M., Yoshida, Y., 2018. Spectral normalization for generative adversarial networks. arXiv preprint arXiv:1802.05957.

Nijkamp, J., Pos, F.J., Nuver, T.T., de Jong, R., Remeijer, P., Sonke, J.-J., Lebesque, J.V., 2008. Adaptive radiotherapy for prostate cancer using kilovoltage cone-beam computed tomography: First clinical results. Int. J. Radiat. Oncol. Biol. Phys. 70 (1), 75–82. http://dx.doi.org/10.1016/j.ijrobp.2007.05.046.

Nijskens, L., van den Berg, C.A., Verhoeff, J.J., Maspero, M., 2023. Exploring contrast generalisation in deep learning-based brain MRI-to-CT synthesis. Phys. Medica 112, 102642. http://dx.doi.org/10.1016/j.ejmp.2023.102642.

Olberg, S., Zhang, H., Kennedy, W.R., Chun, J., Rodriguez, V., Zoberi, I., Thomas, M.A., Kim, J.S., Mutic, S., Green, O.L., Park, J.C., 2019. Synthetic CT reconstruction using a deep spatial pyramid convolutional framework for MR-only breast radiotherapy. Med. Phys. 46 (9), 4135–4147. http://dx.doi.org/10.1002/mp.13716.

Pappas, E.P., Alshanqity, M., Moutsatsos, A., Lababidi, H., Alsafi, K., Georgiou, K., Karaiskos, P., Georgiou, E., 2017. MRI-related geometric distortions in stereotactic radiotherapy treatment planning: Evaluation and dosimetric impact. Technol. Cancer Res. Treat. 16 (6), 1120–1129. http://dx.doi.org/10.1177/1533034617735454.

Peng, J., Shi, C., Laugeman, E., Hu, W., Zhang, Z., Mutic, S., Cai, B., 2020. Implementation of the structural similarity (SSIM) index as a quantitative evaluation tool for dose distribution error detection. Med. Phys. 47 (4), 1907–1919. http://dx.doi.org/10.1002/mp.14010.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing, Cham, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Schmidt, M.A., Payne, G.S., 2015. Radiotherapy planning using MRI. Phys. Med. Biol. 60 (22), R323. http://dx.doi.org/10.1088/0031-9155/60/22/R323.

Seco, J., Evans, P.M., 2006. Assessing the effect of electron density in photon dose calculations. Med. Phys. 33 (2), 540–552. http://dx.doi.org/10.1118/1.2161407.

Spadea, M.F., Maspero, M., Zaffino, P., Seco, J., 2021. Deep learning based synthetic-CT generation in radiotherapy and PET: a review. Med. Phys. 48 (11), 6537–6566. http://dx.doi.org/10.1002/mp.15150.

Spadea, M.F., Pileggi, G., Zaffino, P., Salome, P., Catana, C., Izquierdo-Garcia, D., Amato, F., Seco, J., 2019. Deep convolution neural network (DCNN) multiplane approach to synthetic CT generation from MR images—Application in brain proton therapy. Int. J. Radiat. Oncol. Biol. Phys. 105 (3), 495–503. http://dx.doi.org/10.1016/j.ijrobp.2019.06.2535.

Spearman, C., 1904. The proof and measurement of association between two things. Am. J. Psychol. 15 (1), 72–101. http://dx.doi.org/10.2307/1412159.

Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L., 2021. How to train your vit? data, augmentation, and regularization in vision transformers. arXiv preprint arXiv:2106.10270.

Sun, B., Jia, S., Jiang, X., Jia, F., 2022. Double U-net CycleGAN for 3D MR to CT image synthesis. Int. J. Comput. Assist. Radiol. Surg. 18 (1), 149–156. http://dx.doi.org/10.1007/s11548-022-02732-x.

Taylor, L., Nitschke, G., 2018. Improving deep learning with generic data augmentation. In: 2018 IEEE Symposium Series on Computational Intelligence. SSCI, IEEE, http://dx.doi.org/10.1109/ssci.2018.8628742.

Texier, B., Hémon, C., Lekieffre, P., Collot, E., Tahri, S., Chourak, H., Dowling, J., Greer, P., Bessieres, I., Acosta, O., et al., 2023. Computed tomography synthesis from magnetic resonance imaging using cycle generative adversarial networks with multicenter learning. Phys. Imaging Radiat. Oncol. 28, 100511. http://dx.doi.org/10.1016/j.phro.2023.100511.

Thummerer, A., Huijben, E., Terpstra, M., Gurney-Champion, O., Afonso, M., Pai, S., Koopmans, P., van Eijnatten, M., Perko, Z., Maspero, M., 2023b. SynthRAD2023 challenge design: Synthesizing computed tomography for radiotherapy. http://dx.doi.org/10.5281/zenodo.7781049.

Thummerer, A., van der Bijl, E., Galapon, Jr., A., Verhoeff, J.J.C., Langendijk, J.A., Both, S., van den Berg, C.A.T., Maspero, M., 2023a. SynthRAD2023 grand challenge dataset: Generating synthetic CT for radiotherapy. Med. Phys. 50, 4664–4674. http://dx.doi.org/10.1002/mp.16529.

van Stralen, M., van der Kolk, B.Y.M., Zijlstra, F., Florkow, M.C., Oost, E., Slotman, J., van Osch, J.A.C., Podlogar, M., Hendrikse, J., de Jong, P., Castelein, R.M., Viergever, M.A., Maas, M., Boomsma, M.F., Seevinck, P.R., 2019. BoneMRI of the cervical spine: Deep learning-based radiodensity contrast generation for selective visualization of osseous structures. In: ISMRM 27th Annual Meeting, Montreal, Canada.

Vandewinckele, L., Claessens, M., Dinkla, A., Brouwer, C., Crijns, W., Verellen, D., van Elmpt, W., 2020. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother. Oncol. 153, 55–66. http://dx.doi.org/10.1016/j.radonc.2020.09.008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is All you Need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Vol. 30, Curran Associates, Inc., URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wiesenfarth, M., Reinke, A., Landman, B.A., Eisenmann, M., Saiz, L.A., Cardoso, M.J., Maier-Hein, L., Kopp-Schneider, A., 2021. Methods and open-source toolkit for analyzing and visualizing challenge results. Sci. Rep. 11 (1), 2369. http://dx.doi.org/10.1038/s41598-021-82017-6.

Wieser, H.-P., Cisternas, E., Wahl, N., Ulrich, S., Stadler, A., Mescher, H., Müller, L.-R., Klinge, T., Gabrys, H., Burigo, L., Mairani, A., Ecker, S., Ackermann, B., Ellerbrock, M., Parodi, K., Jäkel, O., Bangert, M., 2017. Development of the open-source dose calculation and optimization toolkit matRad. Med. Phys. 44 (6), 2556–2568. http://dx.doi.org/10.1002/mp.12251.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biom. Bull. 1 (6), 80–83. http://dx.doi.org/10.2307/3001968.

Yang, C., Xu, J., De Mello, S., Crowley, E.J., Wang, X., 2022. GPViT: A high resolution non-hierarchical vision transformer with group propagation. arXiv preprint arXiv:2212.06795.

Zhang, Z., Liu, Q., Wang, Y., 2018. Road extraction by deep residual U-net. IEEE Geosci. Remote Sens. Lett. 15 (5), 749–753. http://dx.doi.org/10.1109/LGRS.2018.2802944.

Zhong, L., Chen, Z., Shu, H., Zheng, K., Li, Y., Chen, W., Wu, Y., Ma, J., Feng, Q., Yang, W., 2023. Multi-scale tokens-aware transformer network for multi-region and multi-sequence MR-to-CT synthesis in a single model. IEEE Trans. Med. Imaging 1. http://dx.doi.org/10.1109/TMI.2023.3321064.

Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J., 2018. Unet++: A nested u-net architecture for medical image segmentation. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer, pp. 3–11. http://dx.doi.org/10.1007/978-3-030-00889-5_1.

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A., 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. ICCV.