Medical Image Analysis (2024)

Contents lists available at ScienceDirect

Medical Image Analysis

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

# Generating Synthetic Computed Tomography for Radiotherapy: SynthRAD2023 Challenge Report - Supplementary Document A

Evi M. C. Huijben[a,1,2], Maarten L. Terpstra[b,c,1,2], Arthur Jr. Galapon[d,2], Suraj Pai[e,2], Adrian Thummerer[d,f,2], Peter Koopmans[g,2], Manya Afonso[h,2], Maureen van Eijnatten[a,2], Oliver Gurney-Champion[i,j,2], Zeli Chen[k], Yiwen Zhang[k], Kaiyi Zheng[k], Chuanpu Li[k], Haowen Pang[l], Chuyang Ye[l], Runqi Wang[m], Tao Song[n], Fuxin Fan[o], Jingna Qiu[o], Yixing Huang[o], Juhyung Ha[p], Jong Sung Park[p], Alexandra Alain-Beaudoin[q], Silvain Bériault[q], Pengxin Yu[r], Hongbin Guo[s], Zhanyao Huang[s], Gengwan Li[t], Xueru Zhang[t], Yubo Fan[u], Han Liu[u], Bowen Xin[v], Aaron Nicolson[v], Lujia Zhong[w], Zhiwei Deng[w], Gustav Müller-Franzes[x], Firas Khader[x], Xia Li[y], Ye Zhang[y], Cédric Hémon[z], Valentin Boussot[z], Zhihao Zhang[aa], Long Wang[aa], Lu Bai[ab], Shaobin Wang[ab], Derk Mus[ac], Bram Kooiman[ac], Chelsea A. H. Sargeant[ad], Edward G. A. Henderson[ad], Satoshi Kondo[ae], Satoshi Kasai[af], Reza Karimzadeh[ag], Bulat Ibragimov[ag], Thomas Helfer[ah], Jessica Dafflon[ai,aj], Zijie Chen[ak], Enpei Wang[ak], Zoltan Perko[al,2], Matteo Maspero[b,c,2,*]

[a] *Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands*
[b] *Radiotherapy Department, University Medical Center Utrecht, Utrecht, The Netherlands*
[c] *Computational Imaging Group for MR Diagnostics & Therapy, University Medical Center Utrecht, Utrecht, The Netherlands*
[d] *Department of Radiation Oncology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands*
[e] *Department of Radiation Oncology (Maastro), GROW School for Oncology, Maastricht University Medical Centre, Maastricht, The Netherlands*
[f] *Department of Radiation Oncology, LMU University Hospital, LMU Munich, Munich, Germany*
[g] *Department of Radiation Oncology, Radboud University Medical Center, Nijmegen, The Netherlands*
[h] *Wageningen University & Research, Wageningen Plant Research, Wageningen, The Netherlands*
[i] *Department of Radiology and Nuclear Medicine, Amsterdam UMC, location University of Amsterdam, Amsterdam, The Netherlands*
[j] *Cancer Center Amsterdam, Imaging and Biomarkers, Amsterdam, The Netherlands*
[k] *School of Biomedical Engineering, Southern Medical University, Guangzhou, China*
[l] *School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing, China*
[m] *School of Biomedical Engineering, ShanghaiTech University, Shanghai, China*
[n] *Fudan University, Shanghai, China*
[o] *Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany*
[p] *Indiana University, Bloomington, USA*
[q] *Advanced Development Engineering, Elekta Ltd, Montreal, Canada*
[r] *Infervision Medical Technology Co., Ltd. Beijing, China*
[s] *Department of Biomedical Engineering, Shantou University, China*
[t] *Independent researchers*
[u] *Department of Computer Science, Vanderbilt University, Nashville, USA*
[v] *Australian e-Health Research Centre, CSIRO, Herston, Queensland, Australia*
[w] *Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California (USC), Los Angeles, California, USA*
[x] *University Hospital Aachen, Aachen, Germany*
[y] *Center for Proton Therapy, Paul Scherrer Institut, Villigen, Switzerland; Department of Computer Science, ETH Zurich, Zurich, Switzerland*
[z] *University Rennes 1, CLCC Eugène Marquis, INSERM, LTSI, Rennes, France*
[aa] *Subtle Medical, Shanghai, China*
[ab] *MedMind Technology Co. Ltd., Beijing, China*
[ac] *MRI Guidance BV, Utrecht, The Netherlands*
[ad] *Division of Cancer Sciences, The University of Manchester, United Kingdom*
[ae] *Muroran Institute of Technology, Hokkaido, Japan*
[af] *Niigata University of Health and Welfare, Niigata, Japan*
[ag] *Image Analysis, Computational Modelling and Geometry, University of Copenhagen, Denmark*
[ah] *IACS, Stony Brook University, NY, USA*
[ai] *Data Science and Sharing Team, Functional Magnetic Resonance Imaging Facility, National Institute of Mental Health, Bethesda, USA*
[aj] *Machine Learning Team, Functional Magnetic Resonance Imaging Facility National Institute of Mental Health, Bethesda, USA*
[ak] *Shenying Medical Technology (Shenzhen) Co., Ltd., Shenzhen, Guangdong, China*

[*] Corresponding author: Heidelberglaan 100, 3508 GA, UMC Utrecht, P.O. Box 85500 Utrecht, The Netherlands, Tel.: +31-88 75 67492;
*e-mail:* m.maspero@umcutrecht.nl (Matteo Maspero)
[1] Equally contributing first authors
[2] Challenge Organizer

[al]*Delft University of Technology, Faculty of Applied Sciences, Department of Radiation Science and Technology, Delft, The Netherlands*

This document is a supplementary document to Huijben and Terpstra et al. "Generating Synthetic Computed Tomography for Radiotherapy: SynthRAD2023 Challenge Report".

## 1. Participant methods

Each subsection briefly describes the methods used by the participating teams. Top five methods were presented in the main paper. The team names correspond to the submission reported on the leaderboard at `https://synthrad2023.grand-challenge.org/evaluation/test/leaderboard`.

### *1.1. ShantouBME (task 1)*

ShantouBME employed a 2D U-net (Ronneberger et al., 2015) for task 1, incorporating an additional convolutional layer in the bottleneck. The model was trained using L1 loss, with separate models for the brain and pelvis. Magnetic resonance (MR) images were normalized at the patient level, while computed tomography (CT) images were normalized using the fixed range of $[-1024, 3000]$ Hounsfield units (HU). Random patches of $224 \times 224$ pixels were sampled to augment the training data. The model processed the full-size 2D slices in one go during testing, and the normalization procedure was reverted. The models were trained for 300 epochs using the Adam optimizer with a learning rate of $3e-4$ and a step descent learning rate scheduler. The epoch with the lowest validation loss was selected for inference.

### *1.2. FGZ Medical Research (task 2)*

FGZ Medical Research employed one collective 2D denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) with a U-Net architecture (Ronneberger et al., 2015) for brain and pelvis data in task 2. The DDPM was conditioned on the cone beam CT (CBCT) image and trained using the mean squared error (MSE) loss. Training used 500 diffusion steps, while inference used 15 diffusion steps, as implemented in denoising diffusion implicit models (DDIM) (Song et al., 2020). Preprocessing involved resizing slices to $256 \times 256$ pixels, clipping intensities to $[-1024, 2000]$ HU, and normalizing to $[-1, 1]$. No data augmentation was applied. The normalization and resizing steps were reversed to produce a synthetic CT (sCT) in HU. The epoch with the lowest validation mean absolute error (MAE) was selected for inference. The model was trained for 200 epochs using AdamW optimizer and a learning rate of $1e-4$ with a warm-up scheduler.

### *1.3. FGH_365 (task 1 & 2)*

Participating in both tasks, team FGH_365 employed a modality-, anatomy- , and site- (MAS)-specific strategy to synthesize sCT images across multiple modalities (tasks 1 and 2), anatomical regions (brain and pelvis), and sites (centers A, B, and C). Their approach consisted of two MAS-specific solutions. For solution #1, separate 3D pix2pix models (Isola et al., 2017) were trained for each of the 11 MAS configurations present in the dataset. Solution #2 consisted of one unified model conditioned on the MAS and was trained on the collective datasets of both tasks. The latter was based on the 3D pix2pix model but incorporated MAS-conditioned dynamic convolution layers (Zhang et al., 2021; Liu et al., 2022a) at the first two and last two layers. For all models, the generator loss included the L1 loss, the adversarial loss (Isola et al., 2017), and an edge-aware loss (Luo et al., 2021; Fan et al., 2023), and the discriminator loss was the binary cross entropy (BCE) loss. MRI intensities were clipped to the $99.5^{\text{th}}$ percentile, and CT images were clipped to $[-1000, 3000]$ HU. All modalities were linearly normalized to $[-1, 1]$ at the patient level. Data augmentation consisted of random patch selection, random affine transformations, flipping, noise addition, and random contrast adjustment. The models of solution #1 were trained on 3D patches of $128 \times 128 \times 128$ voxels for brain and $256 \times 128 \times 64$ voxels for pelvis, while the model of solution #2 was trained on 3D patches of $192 \times 192 \times 128$ voxels. During inference, patches overlapping by 40% (for volumes $\geq 500 \times 280 \times 140$ voxels) or 60% (for volumes $< 500 \times 280 \times 140$ voxels) were selected using a sliding window and overlapping regions were averaged to create the full-size sCT, which was then linearly rescaled from $[-1, 1]$ to $[-1000, 3000]$ HU. FGH_365 proposed an uncertainty-based site prediction algorithm since the acquisition center was unavailable for the test data. This algorithm considered the MAE between the sCTs obtained from solutions #1 and #2 and assumed the correct center to have the lowest MAE. Solutions #1 and #2 were trained for 5000 and 800 epochs, respectively, and both solutions used the AdamW optimizer with a constant learning rate of $2e-4$. The final models were selected based on the epoch with the lowest validation MAE, and the final output was the average of the sCT predictions from the two MAS-specific solutions.

### *1.4. KoalAI (task 1 & 2)*

Team KoalAI used a locally enhanced (LE) generative adversarial network (GAN), training four models for each subtask. The models consisted of a 3D patched-based generator and a mixture of 3D and 2D patch discriminators (Isola et al., 2017). The generator loss consisted of the L1 loss with a weight of 100 and an adversarial loss (MSE) with a weight of 1. The team included different model architectures for the generator, including ResNet (Johnson et al., 2016), UNet (Ronneberger et al., 2015), and DynUNet (Isensee et al., 2019). Two architectures were considered for the discriminator, including a patch

discriminator (PatchD) (Isola et al., 2017) and a LE discriminator (LED) combining a 3D and a 2D PatchD. Specifically, task 1 pelvis used the ResNet generator and LED, and took inputs of $256 \times 256 \times 56$ voxels. The other three subtaks used an ensemble of three model architectures, presented as 'generator & discriminator' in the following. Task 1 brain considered 1) ResNet & LED with an input size of $256 \times 256 \times 56$ voxels, 2) ResNet & PatchD with an input size of $256 \times 56 \times 256$ voxels, and 3) ResNet & LED with an input size of $56 \times 256 \times 256$ voxels. Task 2 pelvis used 1) DynUNet & PatchD with an input size of $128 \times 128 \times 128$ voxels, 2) ResNet & LED with an input size of $256 \times 256 \times 56$ voxels, and 3) UNet & PatchD with an input size of $448 \times 448 \times 64$ voxels. Lastly, task 2 brain used ResNet & LED with three different input sizes of $256 \times 256 \times 56$, 2) $256 \times 56 \times 256$, and 3) $56 \times 256 \times 256$ voxels. MRI data were preprocessed by histogram matching with a random MRI sample, N4 bias field correction, smoothing with a gradient anisotropic diffusion filter and applying the provided body mask. In addition, the arms on the pelvic MRI were removed using a 2D-connected component algorithm. CBCT data were preprocessed with lower-bound intensity scaling (0 to -1024), applying the provided body mask, and clipping intensities to $[-1000, 3000]$ HU. A thresholding algorithm was also applied, followed by denoising using a 2D connected component algorithm to remove bright spots surrounding the body in the pelvis CBCT data. MRI volumes were normalized to $[-1, 1]$ at the patient level, while (CB)CT inputs were normalized to $[-1, 1]$ using the fixed range $[-1024, 3000]$ HU. Data augmentation for both tasks included random patch selection, affine and elastic deformations, random intensity shifts, random contrast adjustments, and random histogram shifts. The models were trained using the Adam optimizer and learning rate $2e-4$. The training was stopped when the validation MAE did not improve for 100 epochs, and the final model was selected based on the best validation MAE. At test time, output volumes were generated from patches with a 25% overlap and averaged using equal weighting. The normalization process was inverted.

### *1.5. USC-LONI (task 1)*

USC-LONI participated in task 1 and employed a 2.5D diffusion model (Ho et al., 2020) followed by two 2D U-Nets (Ronneberger et al., 2015) acting as refinement networks. The diffusion model, which considered multiple axial slices, was trained using the L1 loss and a consistency loss assessing the difference between adjacent slices in the 2.5D data. One refinement network considered axial slices, while the other considered frontal slices, and both were trained using the L1 loss. CT data were normalized to $[-1, 1]$ using a fixed range of $[-1024, 3000]$ HU, and MRI data were normalized at the patient level to $[0, 1]$. The image slices were resized to $256 \times 256$ for the diffusion model, which considered 3 consecutive slices. The original size was used for the 2D refinement networks. No data augmentation was applied. During inference, overlapping 2.5D inputs were selected with a stride of 1, and DDIM sampling (Song et al., 2020) with 20 steps was used for the diffusion model. Overlapping slices were averaged, resizing and normalization steps were reversed, and the two refinement networks were applied to the 2D axial and frontal slices. The diffusion model was trained for 140,000 iterations with a batch size of 16, using a learning rate of $2e{-}4$ for the first 100,000 iterations and $5e{-}5$ for the last 40,000 iterations. The refinement U-nets were trained for 20,000 iterations with batch size 16 and learning rate $2e{-}4$. All models were trained with the Adam optimizer.

### *1.6. UKA (task 1 & 2)*

UKA used a multiplanar approach consisting of three identical 2D U-Net models (Ronneberger et al., 2015) without skip connections. Separate models were trained for each task and anatomical region using 2D axial, sagittal, or coronal slices. The loss function used a masked average of L1 loss and structural similarity index measure (SSIM) calculated at full and half resolution. MRI intensities within the body mask were clipped to the $99^{\text{th}}$ percentile and normalized to $[-1, 1]$ at the patient level. CBCT intensities were first adjusted to be non-negative and CBCT and CT images were clipped to $[0, 3000]$ and $[-1024, 3000]$ HU, respectively. The (CB)CTs were normalized to $[-1, 1]$ using these fixed ranges. To ensure that the input size was a multiple of 8, zero padding was applied, and data augmentation consisted of random flipping. At test time, horizontal and vertical flipping were applied to each input slice, and the predictions were flipped back and averaged with the unaugmented prediction. Finally, the multi-plane 2D predictions were averaged, and the intensities were rescaled to $[-1024, 3000]$ to produce the final 3D sCT. The models were trained using the AdamW optimizer with a $1e-4$ learning rate. An early stopping criterion was employed to terminate the training process when the validation loss did not decrease for five epochs. The epoch with the lowest validation loss was selected for inference.

### *1.7. PSICPT_AI4PT (task 1)*

Team PSICPT_AI4PT participated in task 1 for which they employed a 2.5D patch-based nnU-Net (Isensee et al., 2021) for both regions separately. The models were trained using the L1 loss. MRI inputs were linearly normalized at the patient level, while CT inputs were normalized using the fixed range of $[-1000, 2000]$. For training, 64 sampling planes of $4 \times 128 \times 128$ voxels were randomly sampled for each patient by applying rotation, flipping, and random clipping. At inference time, patches overlapping by $2 \times 64 \times 64$ voxels were selected, and overlapping regions were averaged. Furthermore, the CT normalization procedure was reverted to result in an sCT in HU. The models were trained for 200 epochs using the Adam optimizer. A warm-up and cosine scheduler adjusted the learning rate to $2e-4$.

### *1.8. Breizh-CT (task 1 & 2)*

BreizhCT participated in both tasks using a pix2pix model (Isola et al., 2017) with a six-block residual network (He et al., 2016) generator and a patchGAN discriminator (Isola et al., 2017). Tasks 1 and 2 utilized a 2D and 3D patch-based model, respectively, and two separate models were trained for the brain and pelvic regions. The loss function combined the cGAN loss (Mirza and Osindero, 2014) and a custom perceptual loss

(Johnson et al., 2016) using the ConvNext-tiny architecture (Liu et al., 2022c) pre-trained on ImageNet. The perceptual loss consisted of a style term and a content term, and for both tasks, the style term was computed by comparing sCT and CT; for task 2, the content term was also computed by comparing sCT and CBCT. The perceptual loss leveraged paired training data but without direct voxel-wise supervision between sCT and CT to avoid registration inaccuracies. Preprocessing involved adjusting CBCT intensities using histogram matching with the ground CT during training and the CT from the training set with the highest mutual information during testing. (CB)CT intensities were clipped to $[-1024, 3000]$ HU and divided by 1000, and MRI intensities were clipped to [0, 2000] and divided by 1000. No data augmentation was applied. The model input sizes for task 1 were 2D patches of $224 \times 224$ pixels for the pelvis and $168 \times 168$ pixels for the brain, and the input sizes for task 2 were 3D patches of $32 \times 224 \times 224$ voxels for the pelvis and $66 \times 168 \times 168$ for the brain. For training and testing, patches were selected with a stride equal half the patch. Postprocessing involved taking the median of overlapping regions, reverting preprocessing steps, and clipping to $[-1024, 3000]$. The models were trained for 200 epochs using the AdamW optimizer with a constant learning rate of $1e-4$. The epoch with the best validation MAE was selected for inference.

### *1.9. SubtleCT (task 1)*

SubtleCT participated in Task 1, using a custom 2.5D U-Net (Ronneberger et al., 2015) with residual blocks (He et al., 2016) replacing the convolutional blocks. They adopted a two-stage approach where the model was first trained with an enhanced CT (eCT) and then with the ground truth CT. The model used the L1 loss in combination with the SSIM loss, and two identical models were trained, one for the brain and one for the pelvis. The eCT was created by setting the window width/level of the CT to 1000/350 HU. MRIs were normalized using min-max normalization at the patient level, while CT and eCT were normalized using fixed ranges of $[-1024, 3000]$ and $[-150, 850]$ HU, respectively. No data augmentation was applied. Input for the model consisted of five adjacent slices, which were padded to a size of $288 \times 288$ voxels for the brain and $512 \times 512$ voxels for the pelvis. During inference, the normalization and resizing procedures were reversed. Both models were trained for 200 epochs, with the first 100 epochs devoted to the first stage, and the following 100 epochs to the second stage. The Adam optimizer was used with an initial learning rate of $1e-4$ and an adaptive learning rate decay scheduler. The epoch with the best validation peak signal-to-noise ratio (PSNR) was selected for inference.

### *1.10. MedicalMind (task 2)*

MedicalMind implemented two 2D models inspired by multi-scale gradients (MSG)-GAN (Karnewar and Wang, 2020) for task 2: Model-Brain and Model-Pelvis. The generator was a U-Net (Ronneberger et al., 2015) with a ResNet-50 encoder (He et al., 2016) and a decoder that included an AdaIn block (Karras et al., 2019) before consecutive convolutions. The generator predicted an sCT at five different resolutions, and the loss considered MAE and MSE for each resolution. The discriminator was similar to a VGG network (Simonyan and Zisserman, 2014) but considered all five resolutions as input, concatenated the low-resolution inputs with the down-scaled large-resolution features, and used the BCE loss. Preprocessing involved resizing 2D axial slices to $512 \times 512$ voxels for both brain and pelvis, clipping CT intensities to [-1000, 2048] HU, and no normalization was applied. During training, random rotation, scale, shift, and flip operations were used to augment the data, and at test time, the sCT with the largest resolution was used and resized back to the original input size. Intensities outside the body mask were set to -1000 HU, and no other postprocessing steps were applied. Both models were trained for 106 epochs using the Adam optimizer with an initial learning rate of $6e-4$ and a StepLR scheduler.

### *1.11. mriG (task 1)*

Team mriG employed a patch-based 3D U-Net (Ronneberger et al., 2015) for task 1 for both regions separately. A combined L1 and SSIM loss was used for training. For preprocessing, a combination of rigid (for bones) and deformable (for soft tissue) image registration techniques (Klein et al., 2009) was used. Registration for the pelvis cases was guided by bone segmentations (Kuiper et al., 2021), with the individual bones segmented using the method outlined by Liu et al. (2021). MRI inputs were normalized to $[-1, 1]$ on a case basis using the minimum value and the $99^{\text{th}}$ percentile, while CT inputs were first made non-negative and then divided by 3000. The data were also reordered to the canonical orientation. Data augmentation included spatial (random zoom and rotation) and intensity-based (random contrast adjustments) augmentations. Additionally, patches of $96 \times 96 \times 64$ voxels were sampled randomly during training. During inference, a sliding window approach was used with half-overlapping patches in each dimension, and Gaussian weighting was applied to the edges. Furthermore, the CT normalization procedure was reverted to result in an sCT in HU. The models were trained with the AdamW optimizer for 100,000 iterations with a batch size 12. The learning rate started at $1e-4$ and ended at $1e-5$.

### *1.12. RRRocket_Lollies (task 2)*

RRRocket_Lollies employed a multi-channel 2D cycleGAN (Zhu et al., 2017) with an auxiliary fusion network for task 2. The discriminator networks were as described by Zhu et al. (2017) with BCE loss; however, the generator architectures were modified to include U-Net-like long-range skip connections (Ronneberger et al., 2015) between corresponding down- and up-sampling convolutional levels to preserve contextual information. Also, attention gates were added to the skip connections to emphasize salient features propagated forward from earlier in the network Schlemper et al. (2019). An auxiliary fusion network was added onto the cycleGAN to assist in multi-channel recombination. This had an identical architecture to the generators but contained only a single residual block and short-range residual connections across convolutional layers. MSE loss was used for both the generators

and fusion networks. Individual models with identical architectures were trained for each anatomical region. Preprocessing involved resizing ($n_{\text{slices}} \times 448 \times 448$ voxels for the pelvis and $n_{\text{slices}} \times 304 \times 304$ for the brain), clipping, outlier correction to correct high-intensity voxels on the surface of the patient and multi-channel normalization. The CT and CBCT scans were normalized into three channels using windowing to enhance the contrast of anatomical structures. The full width of the image range $[-1024, 3000]$ HU was captured in channel one. In channel two, a contrast setting was used to view soft tissue structures; for the CT, this was $[-150, 150]$ HU and $[-100, 100]$ HU for the pelvis and brain, respectively. An automated peak finder was implemented to set the level to the CBCT soft tissue peak, with a fixed window width of 150 HU or ±100 HU. In the final channel, the CT and CBCT images were clipped to $[600, 3000]$ HU to capture information about the high-density structures. Using min-max normalization, each channel was independently scaled to $[0, 1]$. Postprocessing included reversal of preprocessing steps and multi-channel combination. To recombine the channels, the full-width image (channel one sCT for the pelvis, fusion network output sCT for the brain) underwent modifications based on specific conditions: values within the narrow range ( −150 to 150 HU) were substituted with channel two values, and values > 600 HU were replaced with channel three values. No data augmentation was applied during training. The models were optimized using the Adam optimizer and initial learning rates of $1e-4$ and $2e-4$ for the generators and discriminators, respectively. After 5 epochs, the learning rate was reduced to 80% of the learning rate every 2 epochs for both generator and discriminator. The models were trained for a maximum of 200 epochs; however, early stopping was applied when total generator loss did not improve for 20 epochs. The optimal model is chosen based on image similarity metrics (MAE, PSNR, and SSIM) calculated on train-time validation data.

### *1.13. SKJP (task 1 & 2)*

For both tasks, SKJP employed a 2.5D U-Net (Ronneberger et al., 2015) as the basis of the synthesis network and replaced its encoder by EfficientNet-B7 (Tan and Le, 2019) with multi-slice inputs and single-slice outputs. The same architecture was used for both brain and pelvis data, but separate models were trained using the L1 loss. In task 1, data preprocessing involved histogram normalization and linear scaling of MRI intensities, while in task 2, linear scaling of CBCT intensities was applied. Furthermore, axial slices were cropped or padded to $320 \times 320$ voxels for the brain and $480 \times 640$ voxels for the pelvis, with the model considering three consecutive slices as input. No data augmentation was performed. During training, 32 2.5D input volumes were randomly sampled from each 3D volume. The initial learning rates were set to 1e−3, 5e−4, 1e−4, 5e−5 for task 1 brain, task 1 pelvis, task 2 brain, and task 2 pelvis, respectively. The model was trained for 100 epochs using AdamW optimizer, and the learning rate was decreased at every epoch with cosine annealing. The epoch with the lowest validation loss was used as the final model.

### *1.14. Reza Karimzadeh (task 1)*

Reza Karimzadeh participated in task 1, employing a 3D patch-based pix2pix (Isola et al., 2017) with a Swin UNETR (Hatamizadeh et al., 2021) generator. Two identical models were trained for the brain and pelvis. The training loss consisted of the L1 loss weighted by the ground truth CT, an SSIM loss, and an adversarial loss. Preprocessing involved linear normalization to $[-1, 1]$ at the patient level for MRI data and at the population level for CT data. During training, random patches of $64 \times 64 \times 64$ voxels were sampled and augmented using random rotations. During inference, patches with 50% overlap were selected, and the final result was obtained by averaging predictions from overlapping regions. During postprocessing, the normalization procedure was reversed to obtain sCT outputs in HU. The model was trained for 1000 epochs using the AdamW optimizer and the fixed learning rate of $1e-5$. The epoch with the best validation loss was used as the final model for inference.

### *1.15. thomashelfer (task 1)*

Team thomashelfer only participated in task 1, where they used different models for brain and pelvis. For the brain, they employed a 3D latent diffusion model (LDM) (Esser et al., 2021; Ho et al., 2020), combining an autoencoder consisting of a U-Net (Ronneberger et al., 2015) generator with a d diffusion model trained on the latent space of the autoencoder. In addition, ControlNet (Zhang et al., 2023) was used to ensure the generation of CT images conditioned on the MRI images. The LDM and ControlNet were implemented using MONAI generative models (Pinaya et al., 2023). The autoencoder was trained with a combination of L1 loss, perceptual loss (Zhang et al., 2018), a patch-based adversarial objective (Rombach et al., 2022), and a KL regularization of the latent space. The diffusion model and ControlNet were trained using the losses suggested by Pinaya et al. (2023). The input data was center-cropped to $192 \times 192 \times 192$ voxels. MRI were normalized by dividing by 3000, and CT images by subtracting −1024 and then by 4024. No data augmentation was applied. Input images were encoded into a latent space of $48 \times 48 \times 48$ voxels. At test time, sCT intensities were rescaled to the original range, and no other postprocessing was applied. The autoencoder was trained for 1000 epochs using Adam optimizer with learning rates of $5e-5$ and $1e-4$ for the generator and discriminator, respectively. The diffusion model was trained for 1000 epochs using AdamW optimizer with a learning rate of $2.5e-5$, and the ControlNet was trained for 500 epochs using Adam optimizer with a learning rate of $2.5e-5$. The epochs with the best validation loss were used at test time.

For the pelvis, the team employed a 3D (patch-based) pix2pix (Isola et al., 2017) with a U-Net (Ronneberger et al., 2015) generator. The generator loss consisted of the L1 loss with a weight of 100 and the adversarial loss (BCE) with a weight of 1. The pelvis data underwent normalization like the brain data, and no data augmentation was applied. The model processed half of the 3D input volume at a time, allowing for varying input sizes, and combined the two halves without overlap. Intensities were rescaled to the original intensity range to produce the final sCT

output. The model was trained for 400 epochs using Adam optimizer with a learning rate of 1e − 3 for both the generator and discriminator. The last epoch was used as the final model for inference.

### *1.16. X-MAN (task 1 & 2)*

X-MAN used a 3D patch-based cGAN (Liu et al., 2022b) with a nine-block ResNet12 (He et al., 2016) generator and a PatchGAN (Isola et al., 2017) discriminator for both tasks. One collective model was trained for each task, combining brain and pelvis patients, using L1 loss and adversarial loss. MRI and CBCT intensities were normalized linearly at the patient level, while CT intensities were not normalized. The resulting sCT intensities were scaled from $[-1, 1]$ to $[-2000, 2000]$ HU before calculating the loss. During training, random patches of $160 \times 160 \times 32$ voxels were selected. Data augmentation included random intensity shifts between $-10\%$ and $10\%$ and random gamma adjustments with gamma ranging from 0.5 to 1.5. At test time, patches overlapping by $32 \times 32 \times 8$ voxels were selected using a sliding window, and overlapping regions were averaged. The models were trained for 100 epochs using the Adam optimizer with initial learning rates set to 2e − 4 and linearly decreasing to zero over all epochs.

## References


Esser, P., Rombach, R., Ommer, B., 2021. Taming transformers for high-resolution image synthesis, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12873–12883.

Fan, Y., Khan, M.M., Liu, H., Noble, J.H., Labadie, R.F., Dawant, B.M., 2023. Temporal bone ct synthesis for mr-only cochlear implant preoperative planning, in: Medical Imaging 2023: Image-Guided Procedures, Robotic Interventions, and Modeling, SPIE. pp. 358–363.

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.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Ho, J., Jain, A., Abbeel, P., 2020. Denoising Diffusion Probabilistic Models. arXiv preprint arxiv:2006.11239 .

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. Nature methods 18, 203–211.

Isensee, F., Jäger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2019. Automated design of deep learning methods for biomedical image segmentation. arXiv preprint arXiv:1904.08128 .

Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A., 2017. Image-to-Image Translation with Conditional Adversarial Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 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.

Karnewar, A., Wang, O., 2020. Msg-gan: Multi-scale gradients for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Karras, T., Laine, S., Aila, T., 2019. A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2009. Elastix: a toolbox for intensity-based medical image registration. IEEE Transactions on Medical Imaging 29, 196–205. doi:10.1109/TMI.2009.2035616.

Kuiper, R.J., Van Stralen, M., Sakkers, R.J., Bergmans, R.H., Zijlstra, F., Viergever, M.A., Weinans, H., Seevinck, P.R., 2021. CT to MR registration of complex deformations in the knee joint through dual quaternion interpolation of rigid transforms. Physics in Medicine & Biology 66, 175024.

Liu, H., Fan, Y., Li, H., Wang, J., Hu, D., Cui, C., Lee, H.H., Zhang, H., Oguz, I., 2022a. Moddrop++: A dynamic filter network with intra-subject co-training for multiple sclerosis lesion segmentation with missing modalities, in: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2022, Springer. pp. 444–453.

Liu, H., Sigona, M.K., Manuel, T.J., Chen, L.M., Caskey, C.F., Dawant, B.M., 2022b. Synthetic CT skull generation for transcranial MR imaging–guided focused ultrasound interventions with conditional adversarial networks, in: Medical Imaging 2022: Image-Guided Procedures, Robotic Interventions, and Modeling, SPIE. pp. 135–143. doi:10.1117/12.2612946.

Liu, P., Han, H., Du, Y., Zhu, H., Li, Y., Gu, F., Xiao, H., Li, J., Zhao, C., Xiao, L., Wu, X., Zhou, S.K., 2021. Deep learning to segment pelvic bones: large-scale CT datasets and baseline models. International Journal of Computer Assisted Radiology and Surgery 16, 749–756.

Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022c. A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11976–11986.

Luo, Y., Nie, D., Zhan, B., Li, Z., Wu, X., Zhou, J., Wang, Y., Shen, D., 2021. Edge-preserving mri image synthesis via adversarial network with iterative multi-scale fusion. Neurocomputing 452, 63–77.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 .

Pinaya, W.H.L., Graham, M.S., Kerfoot, E., Tudosiu, P.D., Dafflon, J., Fernandez, V., Sanchez, P., Wolleb, J., da Costa, P.F., Patel, A., Chung, H., Zhao, C., Peng, W., Liu, Z., Mei, X., Lucena, O., Ye, J.C., Tsaftaris, S.A., Dogra, P., Feng, A., Modat, M., Nachev, P., Ourselin, S., Cardoso, M.J., 2023. Generative ai for medical imaging: extending the monai framework. arXiv preprint arXiv:2307.15208 .

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B., 2022. High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695.

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.

Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D., 2019. Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis 53, 197–207. doi:10.1016/j.media.2019.01.012.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Song, J., Meng, C., Ermon, S., 2020. Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 .

Tan, M., Le, Q., 2019. Efficientnet: Rethinking model scaling for convolutional neural networks, in: International conference on machine learning, PMLR. pp. 6105–6114.

Zhang, J., Xie, Y., Xia, Y., Shen, C., 2021. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1195–1204.

Zhang, L., Rao, A., Agrawala, M., 2023. Adding conditional control to text-to-image diffusion models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 3836–3847.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

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).