Why Registration Quality Matters: Enhancing sCT Synthesis with IMPACT-Based Registration

Valentin Boussot[1][0009−0003−2465−5458], Cédric Hémon[1][0009−0003−6669−5108], Jean-Claude Nunes[1], and Jean-Louis Dillenseger[1]

Univ Rennes 1, CLCC Eugène Marquis, INSERM, LTSI - UMR 1099, F-35000 Rennes, France
https://ltsi.univ-rennes.fr/

Abstract. We participated in the SynthRAD2025 challenge (Tasks 1 and 2) with a unified pipeline for synthetic CT (sCT) generation from MRI and CBCT, implemented using the KonfAI framework [1]. Our model is a 2.5D U-Net++ with a ResNet-34 encoder, trained jointly across anatomical regions and fine-tuned per region. The loss function combined pixel-wise L1 loss with perceptual losses derived from VGG-19, SAM, and TotalSegmentator to enhance structural fidelity. Training was performed using AdamW (initial LR = 0.001, halved every 25k steps) on patch-based, normalized, body-masked inputs (320×320 for MRI, 256×256 for CBCT), with random flipping as the only augmentation. No post-processing was applied. Final predictions leveraged test-time augmentation and five-fold ensembling. The best model was selected based on validation MAE.

Two registration strategies were evaluated: (i) Elastix with mutual information, consistent with the challenge pipeline, and (ii) IMPACT [2], a feature-based similarity metric leveraging pretrained segmentation networks. On local test sets, IMPACT-based registration yielded improved anatomical alignment and lower MAE compared to Elastix. On the public validation set, models trained with Elastix-aligned data performed better, highlighting the impact of registration bias in the evaluation pipeline.

Keywords: synthetic CT · CBCT · MRI · Image synthesis · Multimodal registration · Perceptual loss

1 Introduction

The generation of synthetic computed tomography (sCT) images from magnetic resonance imaging (MRI) or cone-beam CT (CBCT) supports a wide range of clinical workflows, including MRI-only radiotherapy planning and longitudinal follow-up using low-dose imaging. sCT aims to produce CT-like images with accurate Hounsfield units, closely resembling diagnostic-quality CT, thereby enabling dose calculation, image registration, anatomical segmentation, and treatment planning.

When derived from MRI, sCT eliminates the need for an additional CT scan, thus avoiding exposure to ionizing radiation, a particularly important benefit

for pediatric patients and those requiring repeated imaging. When derived from CBCT, sCT is used to overcome the modality's technical limitations (such as artifacts, noise, and inaccurate Hounsfield units), which are inherent to its low-dose acquisition protocol.

However, producing high-quality sCT images remains challenging due to fundamental differences in acquisition physics, intensity distributions, and anatomical visibility across imaging modalities. This task is further complicated by variability in scanner hardware, acquisition sequences, and imaging protocols, which introduce substantial heterogeneity in the training data. Such variability results in inconsistencies in contrast, resolution, and anatomical representation within the same modality [10]. Deep learning methods, particularly supervised encoder-decoder architectures, have recently shown strong potential for cross-modal image synthesis by learning direct mappings between MRI or CBCT and CT [10,13,9]. When high-quality paired data are available, supervised methods consistently outperform unsupervised approaches, offering superior anatomical accuracy and intensity realism in the synthesized CT images [8].

We followed the prevailing trend by adopting a supervised learning approach for synthetic CT generation, leveraging an encoder–decoder architecture that has shown strong performance in medical image synthesis tasks. A fundamental requirement for such supervised methods is the availability of spatially aligned image pairs. However, in clinical practice, multimodal images are rarely perfectly aligned due to inter-session variations in patient positioning and anatomy [3,12], particularly in abdominal and thoracic regions, which are highly sensitive to breathing, organ motion, and filling effects. Consequently, voxel-level misalignments are common and introduce anatomical inconsistencies into the training data. These alignment errors is especially critical in supervised learning, as it can significantly degrade model performance and lead to anatomically implausible or unrealistic predictions. Achieving accurate multimodal registration remains challenging, especially for MRI, where geometric distortions, variable soft-tissue contrast, and the absence of bone signal limit the effectiveness of conventional intensity-based algorithms such as Elastix [7].

To overcome these challenges, feature-based registration methods such as MIND [5] and IMPACT [2] have been proposed. These approaches compare spatially structured representations rather than raw intensities, enabling more robust and anatomically consistent alignment across modalities. Unlike handcrafted descriptors such as MIND, IMPACT leverages deep semantic features extracted from pretrained segmentation models. By exploiting the spatial Jacobian of the feature maps, IMPACT introduces a differentiable loss that encourages anatomical correspondence even under severe appearance shifts, making it particularly effective in this context.

In this study, we propose a unified deep learning pipeline for synthetic CT generation from MRI and CBCT, developed and evaluated in the context of the SynthRAD2025 challenge [11]. Our main contributions are as follows:

- We investigate the impact of registration quality on supevised image synthesis, comparing traditional Elastix-based alignment with a semantic loss-guided approach leveraging the IMPACT metric.
- We introduce IMPACT-Synth, a novel perceptual loss based on the Segment Anything Model (SAM) [6].

Our results highlight that accurate intermodal registration of the training data is critical for generating structurally faithful synthetic CTs. Misaligned training pairs introduce systematic spatial biases that the model may exploit to artificially boost image similarity metrics. However, this gain comes at the expense of anatomical plausibility, often leading to structurally inconsistent or unrealistic outputs. These findings underscore the importance of proper registration as a prerequisite for reliable cross-modal image synthesis.

2 Method

We participate in both tasks of the SynthRAD2025 challenge: MRI to CT and CBCT to CT synthetic image generation, using a unified pipeline based on supervised training of an encoder–decoder architecture optimized with a combination of L1 loss and perceptual losses.

3 Data

3.1 Pre-alignment process

Two registration approaches were investigated:

- **Baseline registration**: This approach used the Elastix parameter files provided in the official SynthRAD2025 GitHub repository. To reduce the impact of large anatomical mismatches between CT and the secondary modality (MRI or CBCT), a manual curation step was applied. A subset of misaligned image pairs was excluded from the training set, following the same strategy as the one used for test set construction in the challenge. As a result, 101 patients were removed from Task 1 and 50 from Task 2.
- **IMPACT-based registration** [2]: This approach relies on a custom setup using the IMPACT semantic similarity metric. For MRI-to-CT registration, the TotalSegmentator-based model `M730` was used with features extracted from the 7^{th} layer in static mode, combined with the MIND loss. For CBCT-to-CT registration, the same model was applied in Jacobian mode, using features from the 2^{nd} layer. All registrations were performed using a three-level resolution pyramid and a final B-spline grid spacing of 10 mm.

3.2 Preprocessing.

After registration, all images were processed within the patient body mask, with voxels outside the mask set to -1. Preprocessing was modality-specific:

- **CT:** intensities were clipped to a fixed range $[-1024, 3071]$ and linearly normalized to the range $[-1, 1]$.
- **CBCT:** intensities were clipped between the minimum and the 99.5th percentile inside the mask, then normalized to $[-1, 1]$.
- **MRI:** intensities were standardized within the mask using zero mean and unit variance, without any intensity clipping.

3.3 Training and Validation Strategy

For both tasks, we adopted the same methodology based on a nested five-fold cross-validation strategy during the development phase. To simulate a local test set, a subset of 75 images was initially held out. The remaining data were used in an inner five-fold cross-validation to tune hyperparameters and select the optimal model configuration. Once the training strategy was finalized, we retrained five models on the full training set, reintegrating the 75 previously held-out images, using a standard five-fold split. The final submission was obtained by ensembling the predictions of these five models.

4 Model

4.1 Spatial Configuration

We adopted a 2.5D strategy for both tasks, in which each input sample is composed of the target slice stacked with a number of adjacent slices to provide contextual information. For CBCT-to-CT synthesis, we used one slice before and one slice after the target, resulting in a three-channel input. For MR-to-CT synthesis, we used two slices before and two slices after, resulting in a five-channel input.

4.2 Network Architecture

Our synthesis pipeline is built upon an encoder–decoder architecture, specifically a U-Net++ [14] with a ResNet-34 encoder [4]. U-Net++ extends the classical U-Net by introducing nested and dense skip connections, which enhance feature propagation, reduce the semantic gap between encoder and decoder feature maps, and facilitate gradient flow during training.

The encoder is composed of residual blocks from ResNet-34, each consisting of convolutional layers followed by batch normalization and ReLU activations. The decoder includes upsampling layers and dense skip connections from multiple encoder depths, as defined by the U-Net++ architecture. This design enables richer multi-scale feature fusion and supports the learning of more detailed anatomical structures.

The final architecture contains approximately 26.07 million trainable parameters.

4.3 Implementation Details

All experiments were implemented using KonfAI [1], our in-house deep learning framework built on top of PyTorch v2.6.0+cu124. KonfAI offers a modular and fully configurable training pipeline, driven by YAML-based experiment management. It natively supports 2.5D patch-based processing, model ensembling, and test-time augmentation, enabling reproducible and scalable experimentation across tasks.

5 Training

5.1 Training Strategy

The model was trained on paired input–target images. For each fold, we first trained a global model from scratch using random normal weight initialization with a standard deviation (gain) of 0.02, covering all anatomical regions. Subsequently, region-specific models were fine-tuned from the global model using the checkpoint that achieved the best performance in terms of MAE on the validation set. Fine-tuning was performed jointly for the abdomen and thorax (AB+TH), while the head and neck (HN) region was treated independently.

Patch-based training was employed, with a patch size of 256×256 for Task 2 and 320×320 for Task 1, using a batch size of 32. Both the initial training and the fine-tuning phases started with a learning rate of 0.001. A StepLR decay schedule was applied, reducing the learning rate by a factor of 2 every 25,000 steps. Early stopping was applied when the validation performance did not improve for 25,000 consecutive steps. All configurations for Task 1 and Task 2 are publicly available on GitHub: Task 1 and Task 2.

5.2 Data Augmentation

Only one random flip per image was applied during training, refreshed at each epoch, which was sufficient to prevent abrupt overfitting. This minimal augmentation strategy also enabled TTA using flipping at inference time.

5.3 Model Selection

For each configuration, two checkpoints were retained: the model achieving the best performance on the validation set, and the final epoch model. Empirically, the final model often yielded superior results on the held-out test set.

5.4 Inference and Ensembling

To enhance prediction robustness and generalization, we employed test-time augmentation (TTA) based on image flipping. Predictions were computed from flipped inputs, then inverted back to the original orientation before being averaged to produce the final output. This was combined with a cross-validation

ensemble, where predictions from models trained on different folds were averaged to produce the final output. This ensembling strategy reduces variance and mitigates overfitting to individual data splits, leading to more stable and consistent results.

5.5 Loss Functions

All models were trained using the MAE loss as a baseline. In addition, we explored perceptual loss functions inspired by the IMPACT registration framework [2]. These losses aim to incorporate semantic consistency by leveraging features extracted from pretrained segmentation networks. Specifically, we evaluated three types of perceptual losses:

- The original VGG-based perceptual loss, computed from intermediate features of a VGG-19 network pretrained on ImageNet.
- A perceptual loss based on features extracted from the SAM [6] (Segment Anything Model) encoder.

6 Evaluation

6.1 Evaluation Metrics

To evaluate the quality of sCT generation in the SynthRAD2025 challenge, we computed the full suite of image similarity metrics defined by the challenge organizers. These include:

- **Mean Absolute Error (MAE)**: average absolute Hounsfield unit difference between sCT and CT.
- **Peak Signal-to-Noise Ratio (PSNR)**: measures the ratio between the maximum possible image intensity and the MSE error.
- **Multi-Scale Structural Similarity Index (MS-SSIM)**: an advanced version of SSIM computed at multiple image scales, reflecting perceptual image quality.

6.2 Model Selection Strategy

We defined the best model based on comprehensive performance across all evaluation metrics (MAE, PSNR, MS-SSIM). Among candidate checkpoints, the model exhibiting the highest aggregate image similarity was selected as the final submission.

7 Results

Table 1. Comparison of synthesis performance between Baseline and IMPACT registration on Task 1 of the local test set.

Region	Metric	Baseline	IMPACT
AB	MAE	64.89	54.80
	PSNR	29.10	30.97
	MS-SSIM	0.91	0.91
HN	MAE	65.15	70.07
	PSNR	30.20	29.18
	MS-SSIM	0.94	0.95
TH	MAE	60.07	55.97
	PSNR	30.76	31.43
	MS-SSIM	0.94	0.95
Aggregated	MAE	63.37	60.28
	PSNR	30.02	30.53
	MS-SSIM	0.93	0.94

Table 2. Comparison of synthesis performance between Baseline and IMPACT registration on Task 2 of the local test set.

Region	Metric	Baseline	IMPACT
AB	MAE	58.46	49.70
	PSNR	31.33	32.09
	MS-SSIM	0.90	0.91
HN	MAE	60.97	51.97
	PSNR	30.38	31.95
	MS-SSIM	0.94	0.96
TH	MAE	50.40	44.04
	PSNR	31.78	31.43
	MS-SSIM	0.92	0.95
Aggregated	MAE	56.61	48.57
	PSNR	31.16	31.82
	MS-SSIM	0.92	0.94

Table 3. Comparison of synthesis performance between baseline and IMPACT registration on the **Public** validation set.

Metric	Task 1		Task 2	
	Baseline	IMPACT	Baseline	IMPACT
MAE	68.20	75.82	52.87	56.05
PSNR	29.81	28.70	32.36	31.65
SSIM	0.92	0.91	0.96	0.95
Dice	0.72	0.70	0.83	0.82
HD95	8.42	8.89	5.40	5.41

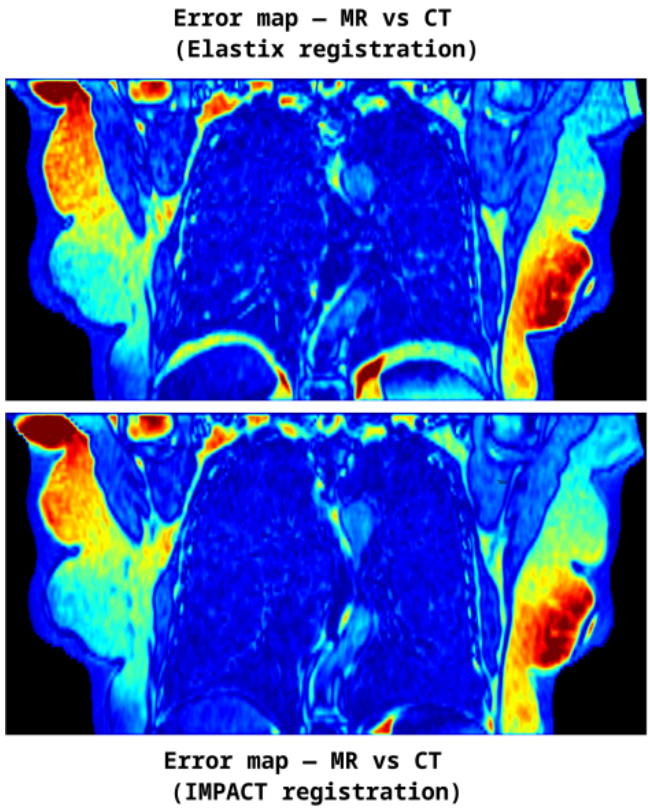

Fig. 1. Error maps for MR/CT registration using Elastix (top) and IMPACT (bottom).

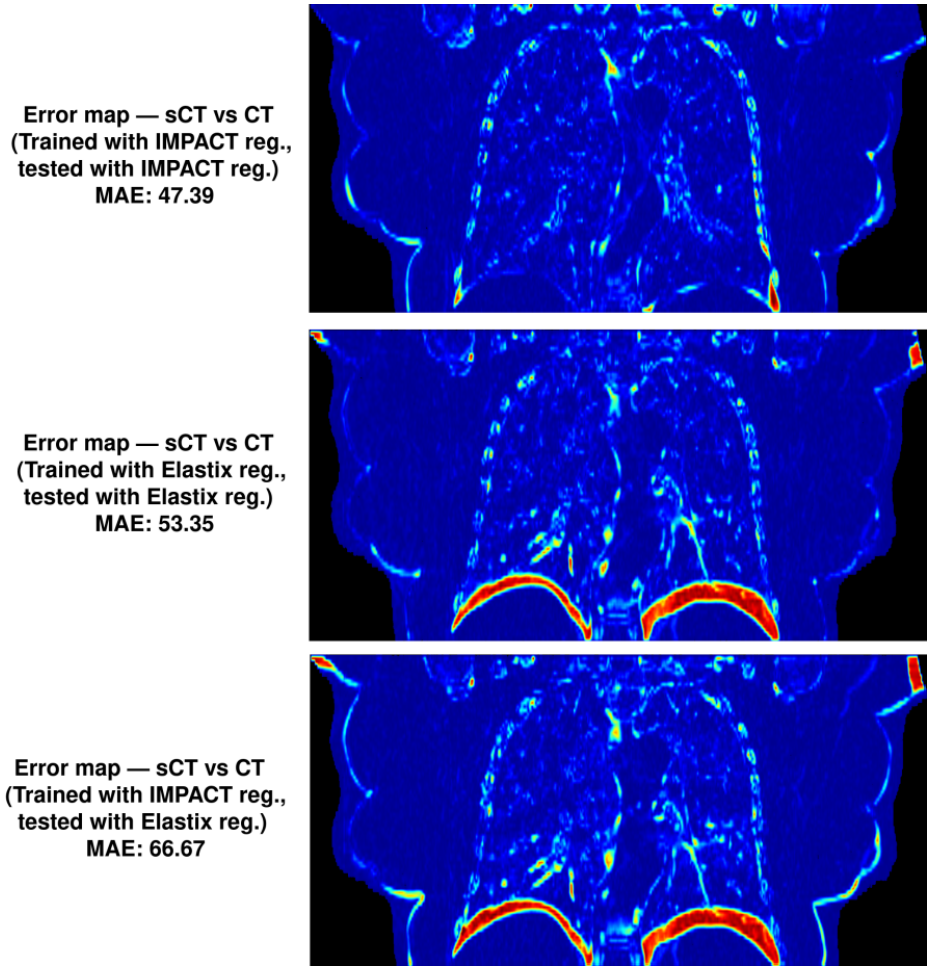

Fig. 2. Error maps of sCT vs CT for different training/test registration combinations.

7.1 Results analysis.

On the local test set, models trained with IMPACT-registered data outperformed those using baseline registration in both Task 1 (Table 1) and Task 2 (Table 2): MAE decreased, PSNR increased, and MS-SSIM improved or remained comparable.

Figure 1 shows that IMPACT achieves better MR/CT alignment. This leads to improved sCT synthesis (Figure 2), with a MAE of 47.39 when training and testing both use IMPACT, compared to 53.35 with baseline registration.

In Task 1, gains were strongest in the Abdomen and Thorax. The Head and Neck showed slightly higher MAE with IMPACT but better MS-SSIM, indicating

more consistent structural alignment. In Task 2, IMPACT outperformed the baseline across all regions and metrics, confirming its robustness.

Qualitative inspection also revealed that baseline-aligned data often introduced anatomical mismatches, especially in the lungs—leading models to "blur" structures and fill air spaces with tissue-like intensities to reduce pixel-wise error. This illustrates how poor registration can bias learning and compromise anatomy.

7.2 Public validation set analysis.

On the public validation set (Table 3), baseline-trained models outperformed those trained with IMPACT. This reflects a distributional bias from the validation pipeline, which favors models using the same registration. In supervised learning, where alignment defines the target, such bias distorts comparisons and does not benefit unsupervised approaches.

Thus, low error can reflect adaptation to registration artifacts rather than true anatomical mapping. This is evident when applying the IMPACT-trained model to Elastix-aligned inputs (MAE = 66.67 vs. 47.39; see Figure 2), where performance collapses due to unseen distortions.

8 Discussion

Our study reveals a central yet often overlooked factor in supervised sCT generation: the quality of intermodal image registration. In supervised learning, voxel-level alignment between input and target modalities is implicitly assumed. Yet in real-world clinical scenarios, such alignment is rarely ensured due to inter-session variability, patient motion, and differences in acquisition protocols. As a result, multimodal images must be pre-aligned using registration algorithms, an essential but imperfect step that may introduce residual errors or spatial biases.

CNN-based encoder–decoder architectures excel at learning complex structural and intensity relationships between modalities. However, this very capacity makes them vulnerable to overfitting to registration artifacts. When exposed to misaligned training pairs, these models may exploit spatial inconsistencies to minimize image similarity losses, leading to deceptively strong metrics (e.g., MAE, PSNR, MS-SSIM) while compromising anatomical realism. This highlights the need to interpret such scores in light of registration fidelity, and to prioritize anatomically meaningful supervision.

IMPACT provides more accurate and anatomically consistent alignments than mutual information (MI) in multimodal registration tasks [2]. In our experiments, these improved registrations appear to positively influence sCT synthesis, leading to better image similarity metrics and more realistic anatomical structures. This supports the idea that registration quality plays an important role in supervised image synthesis.

Beyond registration, we introduce a perceptual loss (IMPACT-Synth) based on a frozen segmentation backbone to enforce structural consistency during syn-

thesis. Complementing standard pixel-wise objectives, this loss leverages semantic cues to guide the model toward anatomically plausible outputs. Visual inspection shows reduced blurring and sharper organ boundaries, underscoring the value of integrating high-level semantic priors into the synthesis process.

IMPACT may therefore be a useful pre-alignment tool in pipelines where training performance depends on good spatial correspondence between modalities. Its ability to handle appearance differences across modalities makes it a practical option for improving synthesis quality in real-world settings.

9 Author contributions

Conceptualization: V. Boussot, C. Hémon. Methodology: V. Boussot, C. Hémon. Software and experiments: V. Boussot. Supervision: J.-L. Dillenseger, J.-C. Nunes. Writing – original draft: V. Boussot, C. Hémon. Writing – review and editing: all authors.

Acknowledgment

The work presented in this article was supported by the Brittany Region through its Allocations de Recherche Doctorale framework and by the French National Research Agency as part of the VATSop project (ANR-20-CE19-0015). Additionally, it was supported by a PhD scholarship Grant from Elekta AB (C.Hémon). The authors have no relevant financial or non-financial interests to disclose. While preparing this work, the authors used 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.

References

1. Boussot, V., Dillenseger, J.L.: Konfai: A modular and fully configurable framework for deep learning in medical imaging (2025). https://doi.org/10.48550/ARXIV.2508.09823, `https://arxiv.org/abs/2508.09823`
2. Boussot, V., Hémon, C., Nunes, J.C., Downling, J., Rouzé, S., Lafond, C., Barateau, A., Dillenseger, J.L.: Impact: A generic semantic loss for multimodal medical image registration. arXiv e-prints pp. arXiv–2503 (2025)
3. Florkow, M., Zijlstra, F., Kerkmeijer, L., Maspero, M., van den Berg, C., van Stralen, M., Seevinck, P.: The impact of mri-ct registration errors on deep learning-based synthetic ct generation. In: Medical Imaging 2019: Image Processing, vol. 10949, pp. 831–7 (2019)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

5. Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady, M., Schnabel, J.A.: Mind: Modality independent neighbourhood descriptor for multi-modal deformable registration. Medical image analysis **16**(7), 1423–1435 (2012)
6. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
7. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE transactions on medical imaging **29**(1), 196–205 (2009)
8. Rossi, M., Cerveri, P.: Comparison of supervised and unsupervised approaches for the generation of synthetic ct from cone-beam ct. Diagnostics **11**(8), 1435 (2021)
9. Sherwani, M., Gopalakrishnan, S.: A systematic literature review: deep learning techniques for synthetic medical image generation and their applications in radiotherapy. Front Radiol **4**, 1385742 (2024). https://doi.org/10.3389/fradi.2024.1385742
10. Spadea, M.F., Maspero, M., Zaffino, P., Seco, J.: Deep learning based synthetic-ct generation in radiotherapy and pet: a review. Medical physics **48**(11), 6537–6566 (2021)
11. Thummerer, A., van der Bijl, E., Galapon, A.J., Kamp, F., Savenije, M., Muijs, C., Aluwini, S., Steenbakkers, R.J., Beuel, S., Intven, M.P., et al.: Synthrad2025 grand challenge dataset: Generating synthetic cts for radiotherapy from head to abdomen. Medical physics **52**(7), e17981 (2025)
12. Yang, H., Sun, J., Carass, A., Zhao, C., Lee, J., Prince, J., Xu, Z.: Unsupervised mr-to-ct synthesis using structure-constrained cyclegan. IEEE T med imaging **39**(12), 4249–4261 (2020)
13. Zhong, Z., Xie, X.: Clinical Applications of Generative Artificial Intelligence in Radiology: Image Translation, Synthesis and Text Generation. Brit J Radiol **1**(1), ubae012 (2024). https://doi.org/10.1093/bjrai/ubae012
14. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018)