Received: 16 September 2023 | Revised: 1 March 2024 | Accepted: 17 March 2024

DOI: 10.1002/mp.17057

RESEARCH ARTICLE

MEDICAL PHYSICS

Enhancing adaptive proton therapy through CBCT images: Synthetic head and neck CT generation based on 3D vision transformers

David Viar-Hernandez[1] | Juan Manuel Molina-Maza[1] | Juan Antonio Vera-Sánchez[2] | Juan Maria Perez-Moreno[2] | Alejandro Mazal[2] | Borja Rodriguez-Vila[1] | Norberto Malpica[1] | Angel Torrado-Carvajal[1]

[1]Universidad Rey Juan Carlos, Medical Image Analysis and Biometry Laboratory, Madrid, Spain

[2]Centro de Protonterapia Quironsalud, Servicio de física médica, Madrid, Spain

Correspondence
David Viar-Hernandez, Universidad Rey Juan Carlos, Medical Image Analysis and Biometry Laboratory, Madrid, Spain.
Email: david.viar@urjc.es

Funding information
Spanish Ministry of Science and Innovation, Grant/Award Number: ID2020-116769RB-I00

Abstract

Background: Proton therapy is a form of radiotherapy commonly used to treat various cancers. Due to its high conformality, minor variations in patient anatomy can lead to significant alterations in dose distribution, making adaptation crucial. While cone-beam computed tomography (CBCT) is a well-established technique for adaptive radiation therapy (ART), it cannot be directly used for adaptive proton therapy (APT) treatments because the stopping power ratio (SPR) cannot be estimated from CBCT images.
Purpose: To address this limitation, Deep Learning methods have been suggested for converting pseudo-CT (pCT) images from CBCT images. In spite of convolutional neural networks (CNNs) have shown consistent improvement in pCT literature, there is still a need for further enhancements to make them suitable for clinical applications.
Methods: The authors introduce the 3D vision transformer (ViT) block, studying its performance at various stages of the proposed architectures. Additionally, they conduct a retrospective analysis of a dataset that includes 259 image pairs from 59 patients who underwent treatment for head and neck cancer. The dataset is partitioned into 80% for training, 10% for validation, and 10% for testing purposes.
Results: The SPR maps obtained from the pCT using the proposed method present an absolute relative error of less than 5% from those computed from the planning CT, thus improving the results of CBCT.
Conclusions: We introduce an enhanced ViT3D architecture for pCT image generation from CBCT images, reducing SPR error within clinical margins for APT workflows. The new method minimizes bias compared to CT-based SPR estimation and dose calculation, signaling a promising direction for future research in this field. However, further research is needed to assess the robustness and generalizability across different medical imaging applications.

KEYWORDS
Adaptive proton therapy, Deep Learning, Pseudo CT synthesis

1 INTRODUCTION

Proton therapy is one of the most promising areas of radiotherapy for several types of cancers, including pediatric, ocular, skull base, and re-irradiated tumors, among others.[1] The benefits of proton therapy are associated with a different radiobiology than photon beams and the reduction of treatment-associated side effects.[2] Similar to conventional radiation therapy, proton therapy requires a virtual simulation based on computed tomography (CT) images to perform dose calculations during treatment planning. For these calculations, it is necessary to convert the Hounsfield Units (HU) obtained in the acquisitions to the stopping power ratio (SPR) of each material.[3]

Additionally, adaptive radiation therapy (ART) plays a vital role in improving treatment outcomes.[4] Its primary objective is to monitor and incorporate any treatment variations that occur during treatment. Such variations can arise from factors such as patient setup, machine delivery deviations, or changes in the patient's anatomy.[5] The significance of adaptation becomes even more critical in proton therapy, where even small deviations in patient anatomy can result in substantial changes in dose due to the precise dose conformality and finite range of proton beams. This underscores the importance of employing adaptive strategies in proton therapy, as highlighted by Paganetti in their work on adaptive techniques.[6]

Cone-beam computed tomography (CBCT) is extensively utilized in ART treatments, playing a crucial role in tasks such as session registration, safety margin adjustment, fraction dose re-calculations, and monitoring the planned dose.[7] However, in adaptive proton therapy (APT) treatments, CBCT can be used for registration and margin adaptation, but it cannot directly estimate the SPR, hindering its utility in dose recalculations.[6]

The inability to directly calculate SPR from CBCT arises from several factors that include higher levels of artifacts (such as streaking and beam hardening), lower image resolution, and limited quantification of information. Furthermore, CBCT also has limitations in differentiating materials and accurately assessing tissue density.

The limitations mentioned above are presented in Figure 1, which provides a visual representation of these deficiencies using a medical phantom.

2 STATE-OF-THE-ART

Several deep-learning methods have been proposed for converting CBCT to pseudo-CT (pCT) images in the last 3 years, mainly focused on improving the accuracy and efficiency of pCT generation while reducing the radiation dose. Some works directly aimed at dose calculations in ART,[8–12] or in APT.[13] The use of a commercial tool for dose calculation has been described recently, reporting an excellent dosimetric agreement between pCT and planning CT.[14]

The initial efforts using deep-learning-based solutions were based on U-Nets.[8,13] Currently, some studies propose the use of DenseNet,[15] while others use conditional generative adversarial networks (cGANs) to generate pCT images.[9–12,16]

Despite the promising results published in recent years, there are still challenges associated with pCT generation, such as the need for large amounts of training data and potential biases in the algorithms.[17] Additionally, there may be limitations in the accuracy of pCT images generated by deep learning models compared to traditional CT images.[18]

Recently, Vision Transformer (ViT) models have gained attention for their ability to capture global contextual information, allowing them to model long-range dependencies in images.[19,20] Unlike Convolutional Neural Networks (CNNs), which process images through convolutional layers and pooling operations, ViTs use the self-attention mechanism developed originally for natural language processing tasks.[21] ViTs are particularly effective for object detection and segmentation and are also flexible and transferable into small datasets.[22] After the first developments, several papers were presented on the combination of visual transformers and convolutions that improve the performance of both.[23–25]

However, as far as the authors know, no previous efforts have proposed the use of 3D ViT in the field of medical image synthesis.

3 MATERIALS

3.1 Dataset

A dataset of 259 paired CBCT-CT images from 59 patients (mean age 30.22 ± 23.04 yo, 28 women) that underwent treatment was retrospectively acquired at Centro de Protonterapia Quironsalud. The inclusion criteria involved the selection of patients with head and neck (H&N) cancer who had both CBCT and CT images available, obtained within a time span of one week.

The CT images were acquired with a General Electric (Chicago, IL, USA) Revolution CT scanner with the following specifications: ASIR-60 iterative reconstruction, 120 kVp, pixel size 0.625 mm, slice thickness 1.250 mm, FOV 50 cm. Furthermore, metal artifact reduction (MAR) filters were applied when needed. Meanwhile, the CBCT images were acquired from equipment installed in the treatment room by IBA with the following specifications: 100 kVp, pixel size 0.5371 mm, slice thickness 1 mm, FOV 30 cm (Louvain-La-Neuve, Belgium).

Data were retrospectively sourced from clinical studies conducted at the center, with ethical approval granted by the local Institutional Review Board and

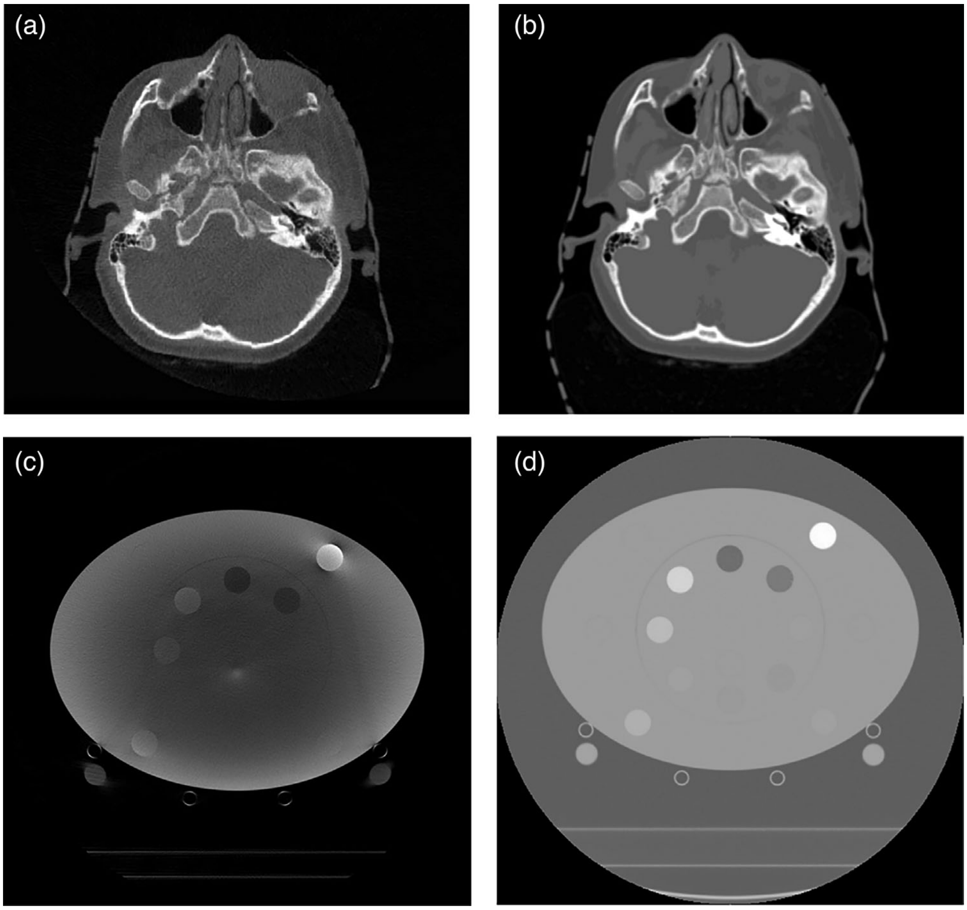

FIGURE 1 Qualitative comparison of the CBCT acquisitions (left column) and the CT acquisitions (right column), using the human head and neck database (a,b) and the medical phantom (c,d). Although CBCT-CT images depict the same phantom/human slice and were acquired with the same kVp, a slight variation arises due to the difference in the field of view (FOV) among them.

strict adherence to the ethical standards delineated in the World Medical Association's Declaration of Helsinki.

A GAMMEX medical phantom (Sun Nuclear, Melbourne, FL, USA) that included 16 tissue substitute inserts was utilized to calibrate the CT scanner and accurately calculate the SPR maps. Additionally, the phantom served as a comprehensive tool for assessing the mentioned limitations of CBCT scanners.

3.2 | Data preprocessing

The image volumes were retrospectively acquired in DICOM format and subsequently converted to NIFTI format immediately before the preprocessing stage. To align the CBCT volumes with the corresponding CT volumes, a rigid registration process was carried out using RayStation Software (RaySearch Laboratories, Stockholm, Sweden) to obtain the transformation matrix.

The transformation matrix obtained from the rigid registration process was employed to project the CBCT onto the CT space, using the 3D Slicer module BRAINs Resample.[26,27] Nearest neighbor interpolation was utilized during this process to ensure that the range values of the registered CBCT images remained unaltered. The default value, corresponding to the CBCT background, was set to −1000.

To ensure consistency and enable straightforward analysis, it is essential to establish a uniform range of values for both image types. To address this variability, a bounding approach was implemented, limiting the pixel values within the range of −1024 to 3072 HU for both CBCT and CT images. Following the determination of the standard range, a min–max normalization technique was implemented to establish the intensity values within the 0–1 range. The upper and lower limits used for the normalization corresponded with the bounded range.

The training process took place on a high-performance NVIDIA cluster comprising eight NVIDIA Ampere A100 GPUs. Despite the high-performance capabilities of the cluster, it was necessary to work with cubic patches of 128 × 128 × 128 voxels to satisfy memory limitations. The complete volume was divided into such patches, applying zero padding to the edges of the input volume edges to include all original voxels into the process. To prevent the presence of visible borders when combining patches for the construction of the final volume, the patches were extracted using a half-overlapping approach. After pCT patches are predicted, the final volume is obtained by cropping the overlapped portions of the patches, obtaining patches of size 64 × 64 × 64, which are joined to recover the pCT volume.

4 | METHODS

CNNs have proven to be highly effective in various computer vision tasks, but they have limitations in capturing global context and long-range dependencies. They primarily rely on local convolutions, which limit their ability to capture global context and long-range dependencies in visual data. This limitation hinders their performance in tasks where global relationships play a significant role. ViT, introduced by Dosovitskiy et al.,[20] addresses this issue by utilizing self-attention mechanisms to model global relationships in an image. By combining the strengths of both architectures, we can achieve superior performance in image processing.

To overcome the aforementioned CNN limitations and take advantage of the ViT architectures, we present a comprehensive study about the inclusion of ViT block in 3D and 2D synthesis tasks. The study is focused on six hierarchical levels of complexity, each corresponding to a specific model. The initial model (Unet-fCNN2D), utilizes 2D convolutions within the U-Net architecture. Moving to the second model (Unet-fCNN3D), we replace the 2D convolutions in the Unet-fCNN2D architecture with more powerful 3D convolutions. Moreover, the third and fourth models, namely Unet-bViT2D and Unet-bViT3D, incorporate 2D and 3D ViT blocks into the bottleneck of Unet-fCNN2D and Unet-fCNN3D architectures, respectively. This progressive incorporation of 2D and 3D ViT blocks further enhances the complexity and capability of the models. Finally, we introduce two additional models, Unet-fViT2D and Unet-fViT3D, whose Unets are constructed exclusively with ViT blocks.

4.1 | Neural network architecture

The U-Net architecture is an autoencoder model used for image generation, which has been modified over the years to improve its performance. The proposed modification (referred to hereafter as Unet) incorporates residual blocks and skip connections to enhance its capabilities. The network consists of three stages: encoding, latent space or bottleneck, and decoding (Figure 2).

The encoding/decoding stages incorporate residual blocks to extract feature maps from the input image or previous stages, reducing/increasing the image size and increasing/reducing the complexity of the maps. Residual blocks include two 3D convolutions, a normalization layer, and a ReLu activation. The number of filters starts at 64 and it is doubled for encoding and halved for decoding stages, where transposed convolutions are used for upsampling. The proposed model relies on three encoding and three decoding stages, and the last convolutional layer is activated using a sigmoid function. Skip connections concatenate feature maps from the encoding and decoding stages, preserving information from multiple scales and complexities, facilitating information flow between stages, and enhancing image generation quality.

Positioned between the encoding and decoding stages, the bottleneck serves as a latent space in the flow of information, encapsulating the essence of the input data in a compact form. This compression of information is important for reducing the dimensionality of feature maps while retaining essential features. Additionally, the use of residual blocks in the bottleneck ensures not only the information compression but also capturing patterns among the feature maps used during the decoding stage. The number of residual blocks in the bottleneck was set at 12 to compensate for the number of parameters in the bottleneck for the ViT models.

4.2 | Vision transformer

The transformer architecture emerged as a powerful tool for natural language processing (NLP), but its potential for image processing tasks, particularly in medical imaging, is currently under investigation.[28] Due to the spatial distribution of information within an image, NLP implementation of transformers needs to be adapted to the image domain. The image or CNN feature map is split into fixed-size patches in a nonoverlapping manner to work as tokens in the same way that words represent tokens in the NLP field. After patch extracting, flattening, and embedding processes, the information has the shape to fit in the self-attention mechanism (Figure 3).

The self-attention mechanism allows the model to consider dependencies between patches across the entire image, facilitating a better understanding of object relationships and semantic hierarchies, and capturing fine-grained details and long-range dependencies between features. In medical imaging, this is particularly important as anatomical structures may be distributed across the image and require a global view for accurate image generation. This enhanced feature representation leads to improved discriminative power and generalization performance.[21] Two different groups of architectures are introduced to harness these advantages in synthesis tasks.

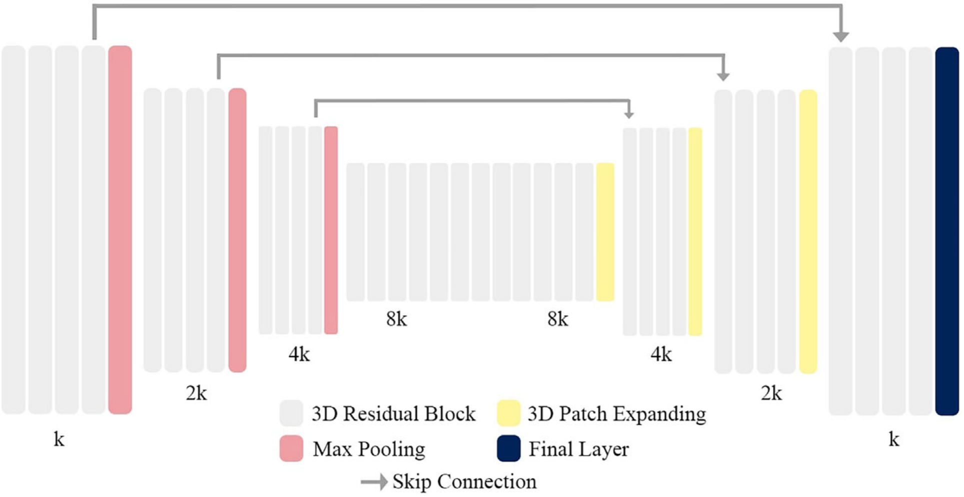

FIGURE 2 The Unet-fCNN3D model is depicted in a schematic representation. This representation is valid also for Unet-fCNN2D with the modification of replacing 3D layers with their 2D counterparts.

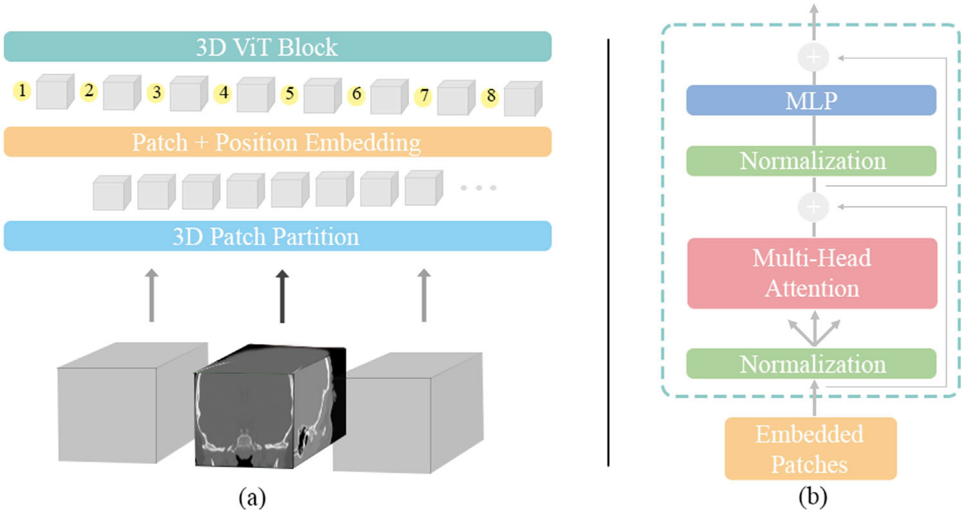

FIGURE 3 The transformer block is shown schematically. (a) Patch partition, patch embedding, and position embedding prepare the patches to get into the transformer block. (b) The transformer block with the appropriate layers.

4.2.1 | ViT blocks in the bottleneck (Unet-bViT)

ViT blocks have demonstrated superior performance when applied to convolved deep feature maps as opposed to the original image space.[29] Therefore, incorporating ViT blocks into the bottleneck not only captures dependencies among complex feature maps but also serves as a potent attention tool for selectively pruning relevant information along the bottleneck.

Within the bottleneck module, the feature maps information that flows from the residual block needs to be patched and projected before it is input into the ViT blocks. The ViT blocks are arranged in a sequence of five successive stages. Every stage has 4 ViT blocks for the 2D model and 3 for the 3D model, that employ a self-attention mechanism. Additionally, residual connections are established between successive stages, thereby enabling the integration of features at various levels. The latent vector shape from the last ViT block needs to be restored to match the feature map shape. Ultimately, the decoder component employs the bottleneck feature maps to progressively reconstruct the synthesized CT image. Figure 4 shows a general overview of the Unet-bViT3D architecture.

In the 2D case, we reshape the input feature map $x \in \mathbb{R}^{HWC}$ into a collection of 2D patches $y \in \mathbb{R}^{N(P^2C)}$, where (H, W) is the shape of the input feature map, C is the number of channels, (P, P) is the shape of each extracted patch, $N = HW/P^2$ is the total number of patches. Regarding the 3D version, the input feature map is $x \in \mathbb{R}^{HWDC}$ and the resultant collection

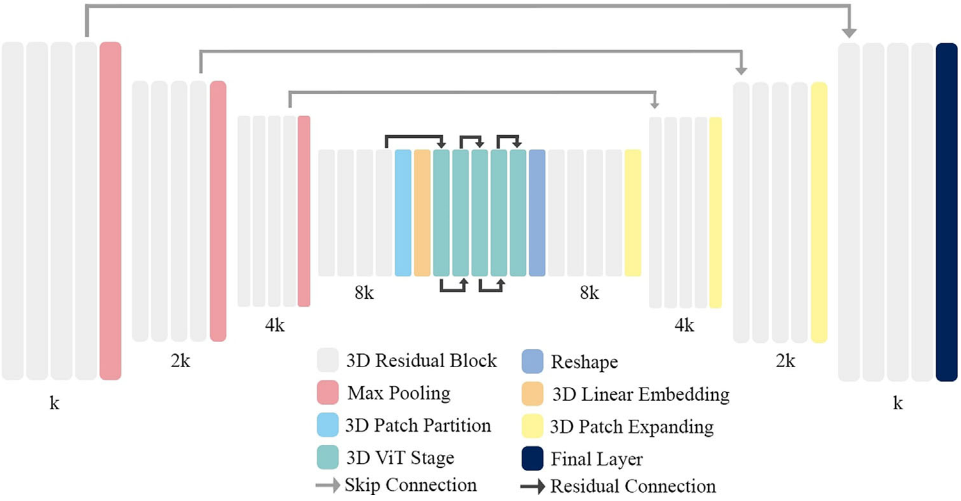

FIGURE 4 The Unet-bViT3D model is depicted in a schematic representation. This representation is valid also for Unet-bViT2D with the modification of replacing 3D layers with their 2D counterparts.

TABLE 1 Summary of the ViT hyperparameters for the four presented models and their amount of parameters.

Models	ViT Blocks	Embedding Dim	Patch/Kernel size	MLP size	Heads/Filters	Params
Unet-fCNN2D	0	–	(3,3)	–	[64,128,256,512]	65M
Unet-fCNN3D	0	–	(3,3,3)	–	[64,128,256,512]	80M
Unet-bViT2D	20	256	(4,4)	512	6	65M
Unet-bViT3D	15	256	(4,4,4)	512	6	75M
Unet-fViT2D	24	[128,256,512,1024]	(4,4)	512	[6,8,12,12]	60M
Unet-fViT3D	16	[128,256,512,1024]	(4,4,4)	512	[6,8,12,12]	75M

of 3D patches is $y \in \mathbb{R}^{N(P^3C)}$, where (H, W, D) is the shape of the input feature map, C is the number of channels, (P, P, P) is the shape of each extracted patch, $N = HWD/P^3$ is the total number of patches. The hiperparameter details can be found in Table 1.

4.2.2 | Full ViT Unet (Unet-fViT)

Other layers that determine the Unet architecture can also be transferred to the ViT structure, allowing for the creation of a pseudo Unet constructed entirely from ViT blocks and transformer tools.

This architecture (Figure 5) relies on three encoder and decoder stages linked one each other by a skip connection operation. Each stage contains three ViT blocks in 2D and two in 3D. After each encoder stage, a patch merging layer reduces the number of patches by half and duplicates the latent vector dimension. Despite ViT transformers being able to capture global patterns over the image, when reducing the number of patches the attention layer is forced to look for different representations of the information and create more sophisticated patterns along the stages.

The bottleneck is made up of four ViT blocks in the 3D model and six ViT blocks in the 2D model, which learn deep feature representations from the encoder outcomes. The feature dimension and resolution remain unchanged in the bottleneck.

Finally, the decoder part restores the original shape of the image through patch-expanding layers and ViT blocks. These patch-expanding layers work in the opposite way of the patch merging, duplicating the number of patches and reducing by half the dimension of the latent vector. Table 1 details the hyperparameters of the models.

4.3 | Experiments

After applying data preprocessing to the original volumes and defining the neural network architecture, the subjects are partitioned into distinct subsets. Out of the 259 pairs of CBCT-CT images, 26 pairs (∼10%) were assigned to the validation set, 25 pairs (∼10%) were assigned to the test set, and the remaining 208 pairs (∼80%) were included in the training set, all the sets relied on a batch size of 3 patches per iteration. All deep

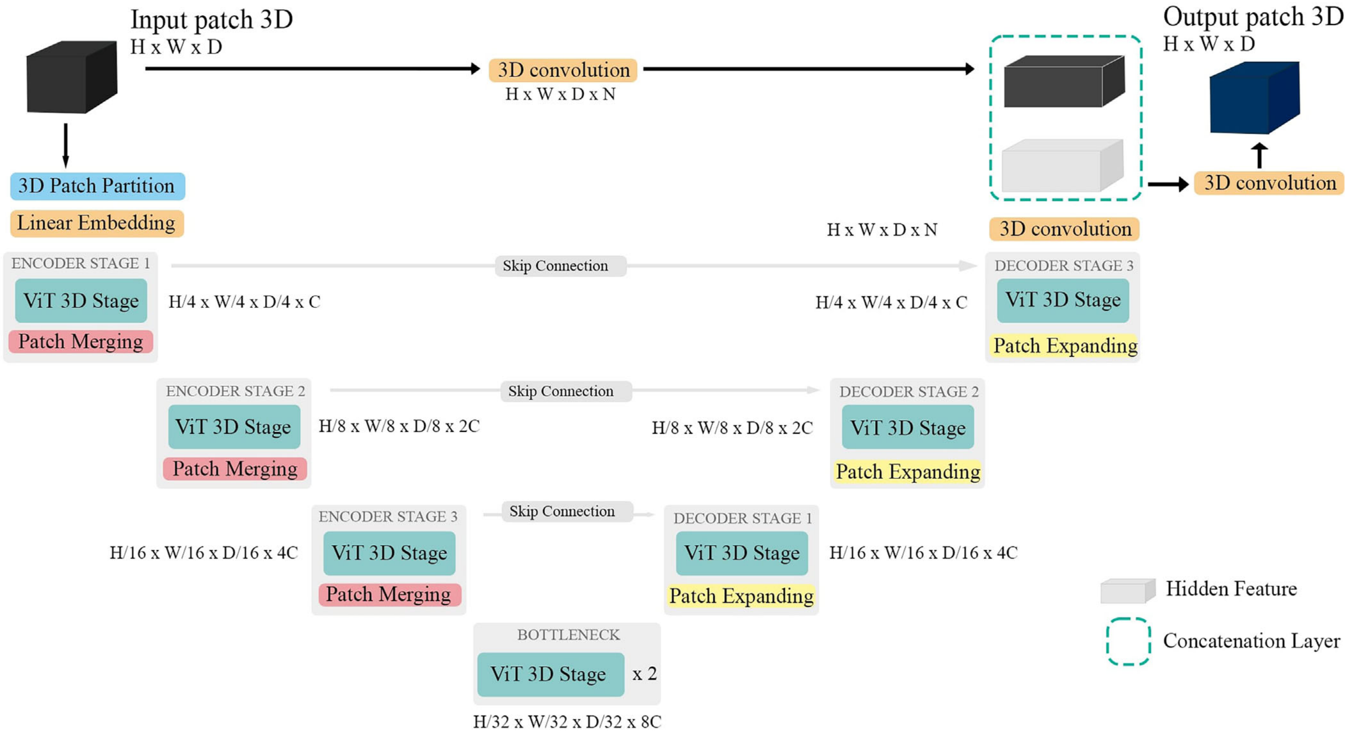

FIGURE 5 The Unet-fViT3D model is illustrated in a schematic representation. Moreover, the architecture described is also employed in Unet-fViT2D, with the distinction that the 3D layers are replaced by their 2D counterparts.

learning models underwent training for 40 epochs, which was determined by an analysis of the validation loss and training process stability. The kernel weights were randomly initialized and Adam was used as the optimizer, with an initial learning rate of 0.0001.

The mean absolute error (MAE) was employed as the loss function for all the models. PSRN, SSIM, and SPR were calculated as evaluation metrics. To calculate the SPR maps, we use the method proposed by Schneider in 1996.[30] It involves calculating the corrected HU and the electron density. The relative percentage error between both SPR maps is then calculated.

The quantitative performance of the various methods was evaluated by comparing the computed metrics between ground truth images and synthetic images. A repeated measures analysis of variance (ANOVA) was employed to assess the impact among the different proposed methods, considering the direct use of CBCT also as a method. Subsequently, paired-samples Wilcoxon signed-rank tests were performed to determine if the performance of each method significantly differed from the other methods. Statistical significance was considered when the *p*-value was below 5%.

Finally, the viability of the generated pCT was clinically assessed with a dose calculation evaluation. This step is not only important to underscore the robustness of the method but also reinforces the potential clinical utility of the generated pCT images.

5 | RESULTS AND EVALUATION

The results demonstrate a promising correlation between the pCT and ground truth CT. Figure 6 shows a representative pCT slice for the six presented methods as well as CBCT and ground truth CT. All the pCTs obtained using our different methods demonstrate a significant improvement when comparing image texture and definition to the CBCT image. Upon visual inspection, the pCT exhibits high quality and our methods prove to be robust, as they can improve the bone contours and differentiate cortical bone from bone marrow, while also improving the differentiation between muscle and fat (Figure 7). ViT architectures seem to accurately characterize the patient-specific tissues, improving the Unet results where the predicted CT images presents more evident noise. Remarkably, the estimated shape of the soft tissue is accurate, despite synthesizing the image from data acquired with a completely different field of view (FOV) and intensity. In terms of image texture, pCT images appear slightly smoother in comparison to the ground-truth CT, although this difference is mitigated in Unet-fViT architectures. Nonetheless, the results highlight the potential and efficacy of the methods in generating accurate pCT images, even with the challenge of noisy CBCT images as a starting point.

A more detailed view of the pCT for the Unet-fViT models, including sagittal, coronal, and axial slices of a representative subject is depicted in Figure 7. As noted

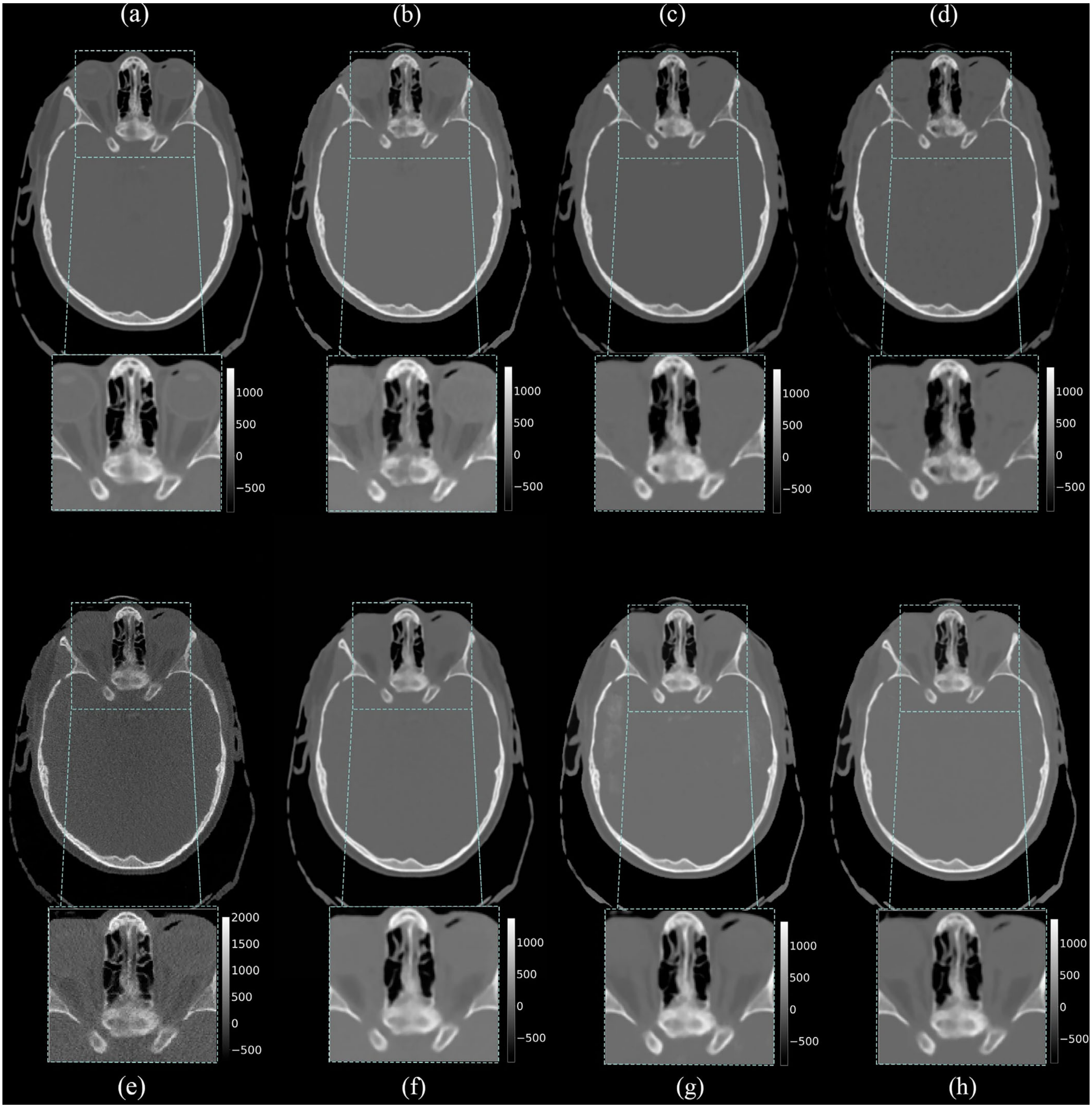

FIGURE 6 Qualitative evaluation of pCT predictions for the six methods: (a) ground truth CT, (b) Unet-fViT3D, (c) Unet-bViT3D, (d) Unet-fCNN3D, (e) CBCT, (f) Unet-fViT2D, (g) Unet-bViT2D, and (h) Unet-fCNN2D.

before, ViT appears to achieve better tissue differentiation, reducing the error in cortical and marrow bones and even in the patient mask prediction. Furthermore, the comparison between 2D and 3D models shows a similar performance for ViT 2D models when we evaluate them in all the image planes (sagittal, coronal, and axial), recognizing the importance of the architecture over the input dimensions in clinical applications.

Figure 8 shows the relative change (RC) maps based on the SPR images calculated using the different images shown in Figure 6 and the one calculated using the ground truth CT. These RC maps demonstrate an even more significant improvement when using the Unet-fViT method, achieving the slightest difference in reference with the result obtained using the ground truth CT. The introduction of ViT blocks in the architecture helped to not only concentrate attention along sensitive regions but also capture the essential information across the image, thus reducing the occurrence of artifacts, particularly beam hardening and

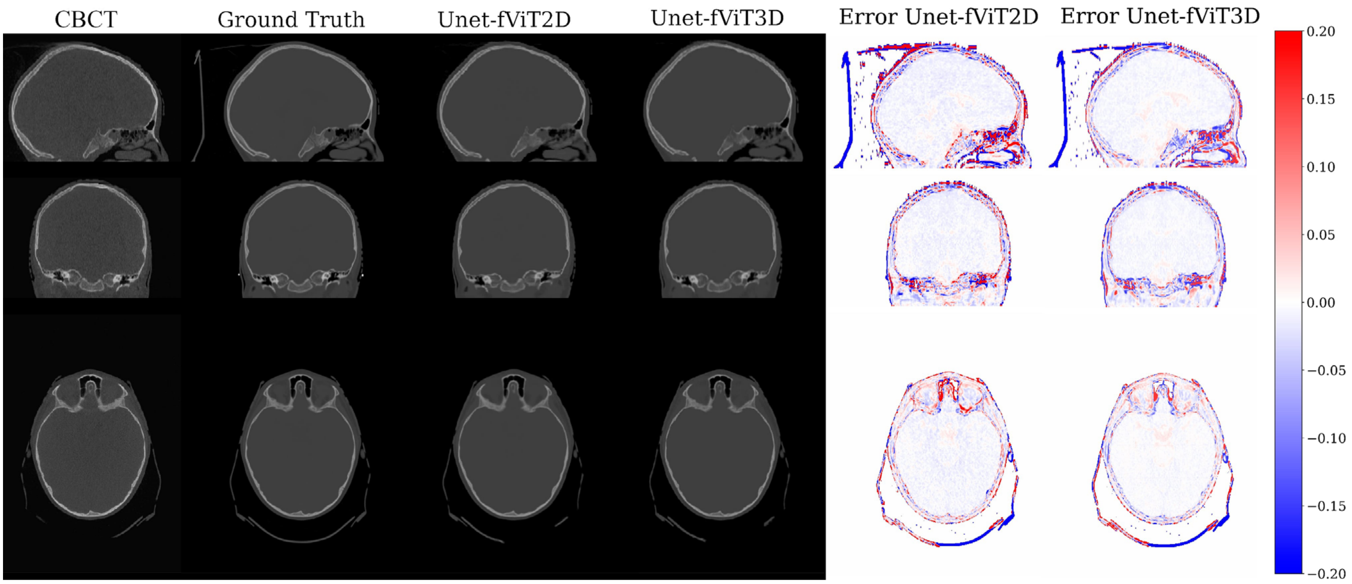

FIGURE 7 Sagittal (top), coronal (middle), and axial (bottom) slices for CBCT, ground truth, pCT using the Unet-fViT2D model and the Unet-fViT3D model, and RC maps based on HU difference between pCT and ground truth.

TABLE 2 Summary of the models' performance and the CBCT metrics presented for comparison.

Models	MAE (HU)	PSNR	SSIM	SPR (%)
CBCT	122.973 ± 21.106	22.861 ± 1.808	0.909 ± 0.021	NA
Unet-fCNN2D	82.186 ± 16.966	26.975 ± 1.490	0.955 ± 0.016	6.311 ± 1.143
Unet-fCNN3D	72.206 ± 15.311	27.652 ± 1.617	0.96 ± 0.014	4.705 ± 1.444
Unet-bViT2D	75.637 ± 16.365	27.695 ± 1.678	0.948 ± 0.015	4.522 ± 1.231
Unet-bViT3D	65.891 ± 17.418	28.249 ± 1.896	0.963 ± 0.016	4.414 ± 1.646
Unet-fViT2D	62.365 ± 18.657	29.664 ± 2.024	0.965 ± 0.016	3.545 ± 1.678
Unet-fViT3D	58.544 ± 17.388	30.251 ± 1.838	0.967 ± 0.014	3.335 ± 1.180

blooming effects (Figure 11). All the models had difficulties in the prediction of certain anatomical areas with insufficient information coming from the CBCT or with deep interpatient variations. The UNet architecture without ViT blocks produced worse results in the SPR calculation as it transferred artifacts into the pCT. The original CBCT was not introduced in the SPR calculation since the calibration process cannot be carried out for devices with the mentioned image quality restrictions.

To quantitatively assess the accuracy of the image-to-image task, we computed several image quality metrics, for all subjects, summarized in Table 2. Violin plots of Figure 9 show how 3D models outperformed 2D models in all the cases. Moreover, the amount of ViT blocks had a positive impact on the model results, achieving better results when the ViT blocks were distributed along the early stages of the model.

All the proposed pCT methods performed better than CBCT for all metrics. ANOVA test revealed a statistically significant effect for the *"model"* in all metrics (PSNR: $F_{5,21} = 10.348$, $p = 2.848$ x 10^{-7}; MAE: $F_{5,21} = 48.433$, $p = 1.523$ x 10^{-24}; SSIM: $F_{5,21} = 23.417$, $p = 1.834$ x 10^{-14}; SPR RC: $F_{5,21} = 30.878$, $p = 7.134$ x 10^{-18}). For all four metrics, the decomposition of such statistical results using paired comparisons by means of the Wilcoxon test revealed that all methods are statistically different from each other ($p's < 0.01$). Such quantitative results are in accordance with the visual results presented before with Unet-fViT3D outperforming the other methods in terms of both image quality and dose calculations. Thus, subsequent analyses exclusively focus on Unet-fViT3D as the chosen method for the analysis.

Additional analyses for the Unet-fViT3D model included a voxel-by-voxel correlation analysis as well as a Bland–Altman analysis between pCT and CT for all participants (Figure 10). There was an excellent correlation between the ground truth CT and the synthetic Unet-fViT3D volume ($m = 0.9696$, adjusted $R^2 = 0.9969$, $\rho = 0.9879$, $p \leq 0.001$). The Bland–Altman plot between

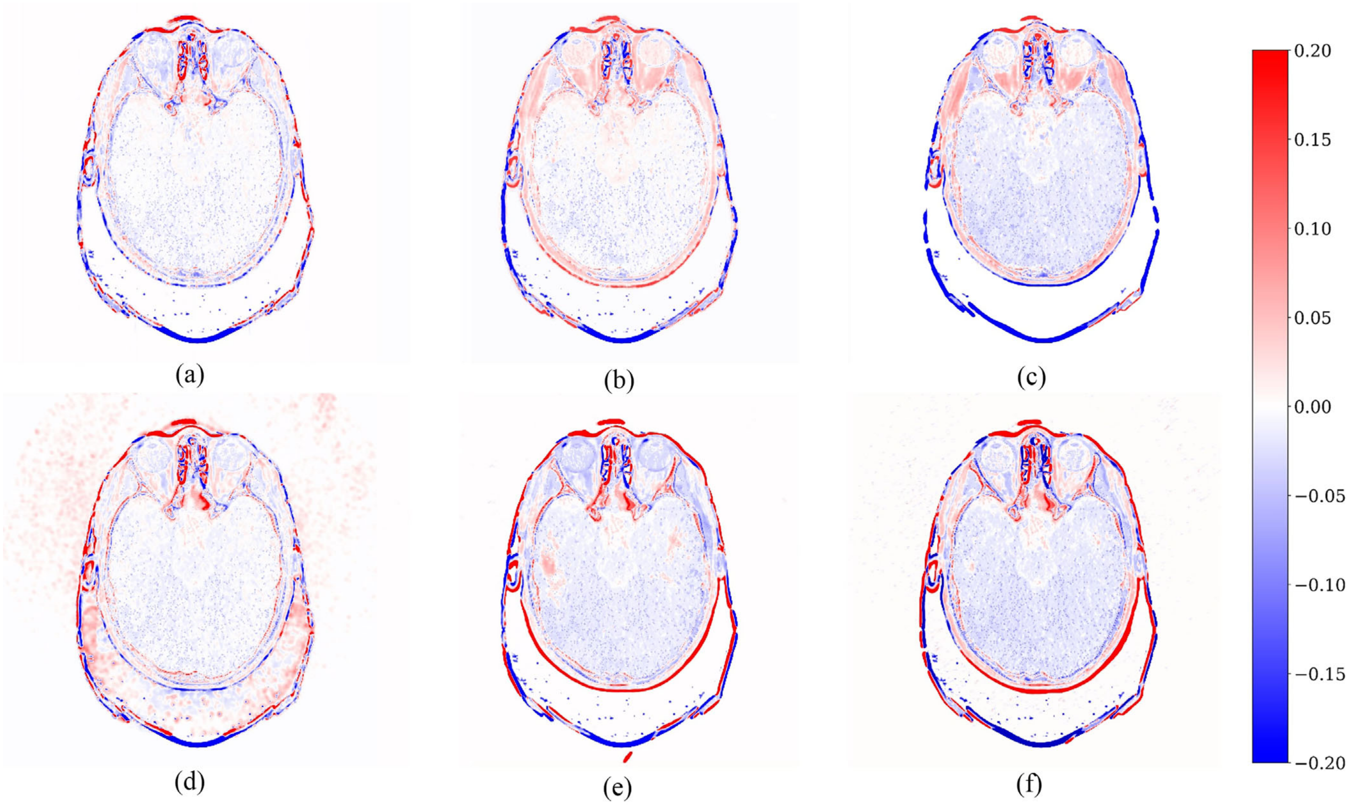

FIGURE 8 Evaluation of RC for SPR in the presented methods: (a) Unet-fViT3D, (b) Unet-bViT3D, (c) Unet-fCNN3D, and (d) Unet-fViT2D, (e) Unet-bViT2D, (f) Unet-fCNN2D.

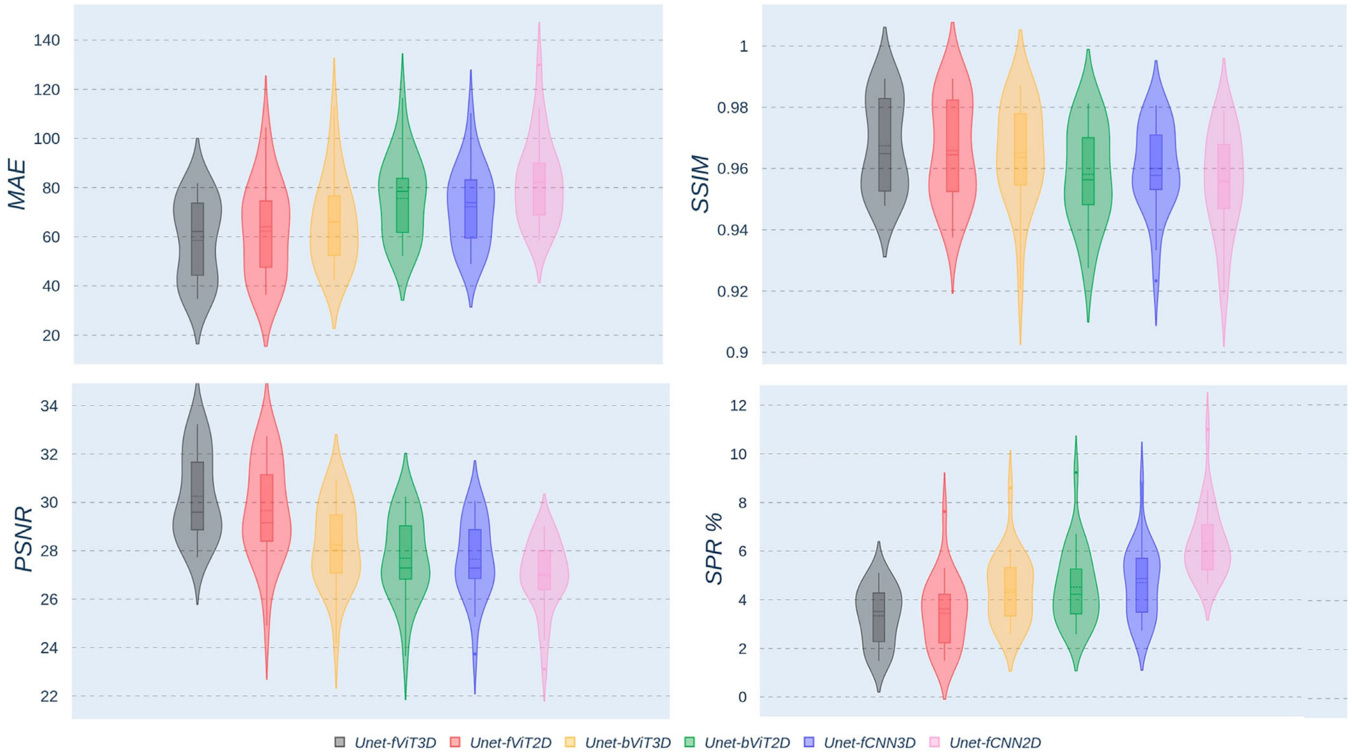

FIGURE 9 Violin plots illustrating the distribution of each evaluation metric.

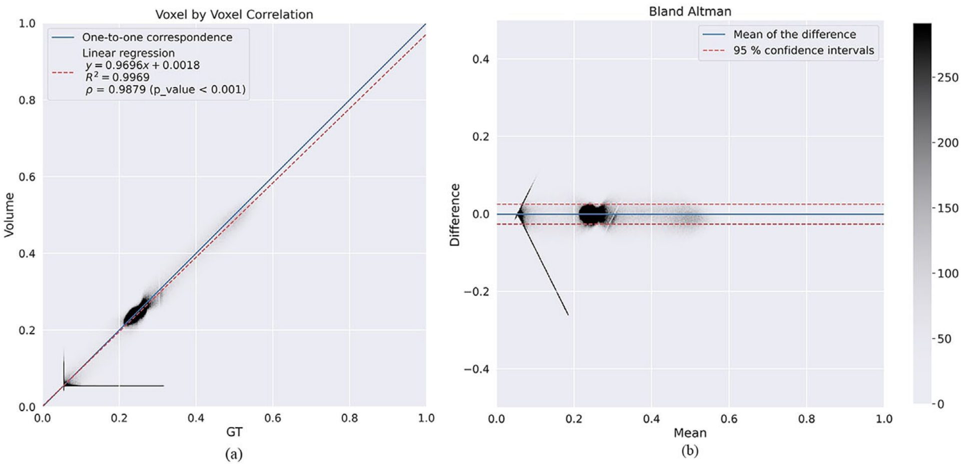

FIGURE 10 Visualization of (a) correlation analysis and (b) Bland Altam analysis using the Unet-fViT3D model for the generated pCT.

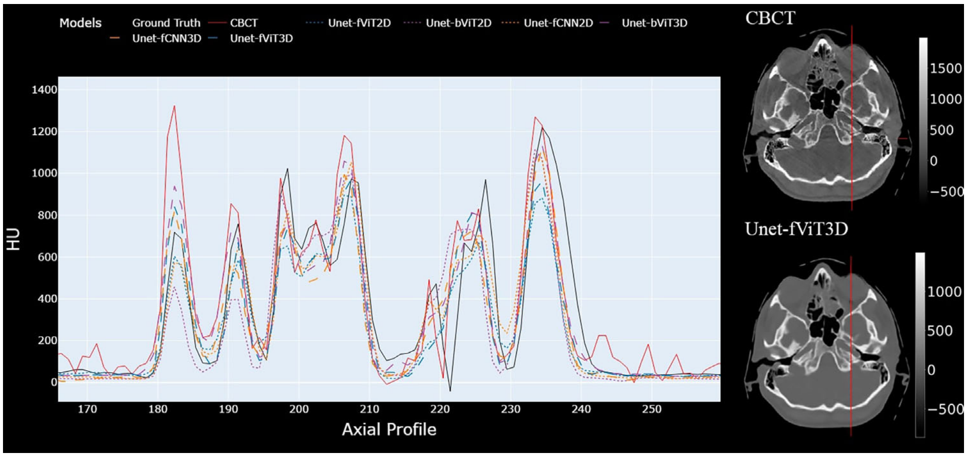

FIGURE 11 Evaluation of the CBCT artifact correction for all the presented models. The axial profile shown in the figure has been zoomed-in due to visualization constraints. The profile studied is signaled with a red line in the presented slices.

the ground truth CT and the synthetic volume showed that the average of differences and variability was low (0.001± 0.05); with 95% confidence interval, with the difference between the proposed method and the ground truth accumulating around the soft tissue areas contained within the skull, which represent the majority of the tissues in the image.

The assessment of artifact transfer (see Figure 11) evaluated from an axial profile line demonstrates the outstanding performance of all ViT models in mitigating CBCT artifacts in the predicted pCT. The Unet-fViT models excelled in providing the most accurate representation of soft tissue, accompanied by a robust correlation in voxel distribution compared to the ground truth. Notably, most of the methods presented a similar performance in high-density tissues.

Concerning dose evaluation, as can be seen in Figure 12d, the histogram curves for the interesting magnitudes (chiasm, blue; right maxillary, purple; lacrimal gland right, yellow; eye right, light blue; CTV 42.5 Gy and CTV 50.4 Gy) present small differences between the control CT that was performed the same day and the pCT coming from the CBCT (of the same day). Also, presented in Table 3, it can be seen the amount of dose for various representative tissues. The tissues labeled with *Plan dose* represent the original CT, whereas the label *Evaluation dose* represents the pCT tissues. Most of the tissues show small differences in the dose, excluding chiasm and maxillary bone, which present higher variations. The dose variation information is illustrated in Figure 12c, which shows the difference in percentage for all the tissues, signaling differences up to 3% in the slices showed in Figure 12a. These dose variations can be attributed to the mean difference in the HU and the impact of this difference in range estimation. Despite these higher variations, the tumor areas were comprised of tissues with minor differences.

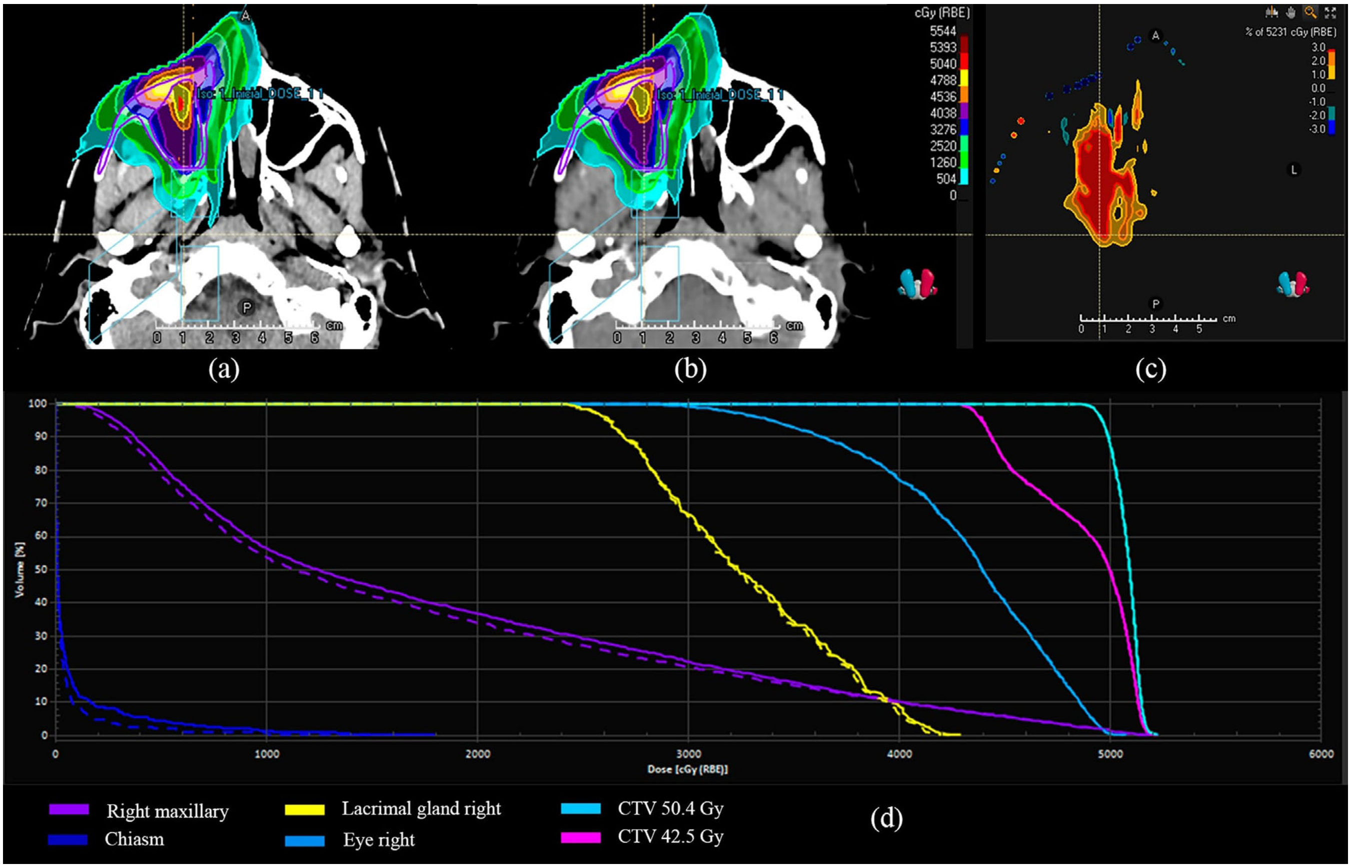

FIGURE 12 Visualization of (a) dose distribution using control CT, (b) dose distribution using pCT with Unet-fViT3D, (c) dose difference in percentage, and (d) dose-volume histogram for the main structures of interest.

TABLE 3 Summary of tissues dose values.

Name	Dose [cGy (RBE)]						
	D99	D98	D95	Average	D50	D2	D1
Plan dose: Body	0	0	0	172	0	3388	4371
Evaluation dose: Body	0	0	0	170	0	3357	4358
Plan dose: Chiasm	0	0	1	74	10	833	1034
Evaluation Dose: Chiasm	0	0	1	45	9	473	600
Plan dose: CTVp_post_50.4 = GTVp_post + 5mmAC	4909	4932	4963	5081	5095	5177	5184
Evaluation Dose: CTVp_post_50.4 - GTVp_post + 5mmAC	4913	4935	4964	5082	5094	5175	5186
Plan dose: Eye_R	3012	3155	3384	4317	4395	4960	4978
Evaluation Dose: Eye_R	3023	3160	3395	4321	4394	4962	4978
Plan dose: Glnd_Lacrimal_R	2485	2544	2628	3290	3233	4148	4174
Evaluation Dose: Glnd_Lacrimal_R	2470	2531	2615	3276	3221	4114	4130
Plan dose: Maxilla_R	155	202	283	1773	1232	4936	5027
Evaluation Dose: Maxilla_R	110	150	235	1690	1131	4930	5021

6 | DISCUSSION

The generation of accurate SPR maps is a basic step in APT. However, while CBCT is a valuable tool that allows for accurate patient positioning and tumor tracking, it cannot be directly applied to APT treatments due to the inability to estimate the SPR from these images. In this context, several CBCT-based pCT approaches have been proposed in the literature providing promising results, but most of them are based on CNNs, restricting the potential use of nonlocal features and relationships found in the images. In this work, we propose the combination of a CNN with one of the latest contributions on ViT to account for these nonlocal contributions to assess the overall benefit of increasing the architecture complexity.

We have evaluated our method both qualitatively and quantitatively considering CT, SPR maps values as well as a dose calculation evaluation. The visual (Figure 6) and quantitative (Table 2) analyses of image quality showed CT and pCT similarity. We compared between Unet-fCNN, Unet-bViT, and Unet-fViT to determine if the ViT blocks were worth it despite offering increased network complexity. Our focus was on the SPR maps and dose calculation obtained from the comparison, and the results indicate that the use of transformers reduces noise and artifacts in the final image, resulting in enhanced feature maps and improved tissue characterization.

Our study presents several limitations. First, our dataset is comprised of a broad age range, combining pediatric and adult patients. Specific scenarios taking into account both separate datasets could be desirable. Second, the 3D patches used to train our model are relatively small in this work due to GPU memory limitations. The nonlocal property of the Unet-fViT3D model could be further exploited by using larger patches when fragmenting the volume, potentially providing better results. Future work will focus on further improving the architecture by using a more complex loss function (e.g., comparing the SPRs of synthetic CT and ground truth CT at each training iteration), including our Unet-fViT3D model as the generator in a GAN architecture, or including magnetic resonance images as input, in addition to CBCT, to improve soft-tissue definition.

The dose evaluation study can be seen as a proof of concept that the method works satisfactorily. However, further evaluation is needed to integrate these algorithms in the APT context.

7 | CONCLUSION

We have outlined an enhanced method for developing and initially validating a ViT 3D architecture approach. This approach is designed to estimate and generate pCT images using standard CBCT images. The introduction of the Unet-fViT3D architecture has significantly reduced the SPR error within the clinically recommended margins for APT workflows. Moreover, it has been demonstrated that transformers excel in synthesis tasks, as they have previously shown in segmentation processes. This new method introduces minimal bias when compared to the current standard, which relies on CT-based approaches for SPR estimation.

ACKNOWLEDGMENTS

This study is part of the project PID2020-116769RB-I00, Adaptive Proton Therapy Using Artificial Intelligence (ADAPT-AI).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

ORCID

David Viar-Hernandez
https://orcid.org/0000-0002-1938-2980
Juan Manuel Molina-Maza
https://orcid.org/0009-0007-7137-2572
Juan Antonio Vera-Sánchez
https://orcid.org/0000-0002-3594-9457
Juan Maria Perez-Moreno
https://orcid.org/0000-0001-6271-1842
Alejandro Mazal
https://orcid.org/0000-0003-1391-8446
Borja Rodriguez-Vila
https://orcid.org/0000-0003-4779-0225
Norberto Malpica
https://orcid.org/0000-0003-4618-7459
Angel Torrado-Carvajal
https://orcid.org/0000-0002-1540-2809

REFERENCES

1. García S, Cabrera P, Herruzo I, et al. Recomendaciones para la SEOR para la protonterapia. SEOR; 2019.
2. Paganetti H, Blakely E, Carabe-Fernandez A, et al. Report of the AAPM TG-256 on the relative biological effectiveness of proton beams in radiation therapy. *Med Phys*. 2019;46:e53-e78.
3. Shafai-Erfani G, Lei Y, Liu Y, et al. MRI-based proton treatment planning for base of skull tumors. *Int J Part Ther*. 2019;6:12-25.
4. Yan D, Vicini F, Wong J, Martinez A. Adaptive radiation therapy. *Phys Med Biol*. 1997;42:123.
5. Sonke J-J, Aznar M, Rasch C. Adaptive radiotherapy for anatomical changes. In: *Seminars in Radiation Oncology*. Vol 29. Elsevier; 2019:245-257.
6. Paganetti H, Botas P, Sharp GC, Winey B. Adaptive proton therapy. *Phys Med Biol*. 2021;66:22TR01.
7. Posiewnik M, Piotrowski T. A review of cone-beam CT applications for adaptive radiotherapy of prostate cancer. *Physica Med*. 2019;59:13-21.
8. Chen L, Liang X, Shen C, Jiang S, Wang J. Synthetic CT generation from CBCT images via deep learning. *Med Phys*. 2020;47:1115-1125.
9. Zhao J, Chen Z, Wang J, et al. MV CBCT-based synthetic CT generation using a deep learning method for rectal cancer adaptive radiotherapy. *Front Oncol*. 2021;11:655325.
10. Gao L, Xie K, Wu X, et al. Generating synthetic CT from low-dose cone-beam CT by using generative adversarial networks for adaptive radiotherapy. *Radiat Oncol*. 2021;16:1-16.
11. O'Hara CJ, Bird D, Al-Qaisieh B, Speight R. Assessment of CBCT-based synthetic CT generation accuracy for adaptive radiotherapy planning. *J Appl Clin Med Phys*. 2022;23:e13737.
12. Wang X, Jian W, Zhang B, et al. Synthetic CT generation from cone-beam CT using deep-learning for breast adaptive radiotherapy. *J Radiat Res Appl Sci*. 2022;15:275-282.
13. Thummerer A, Zaffino P, Meijers A, et al. Comparison of CBCT based synthetic CT methods suitable for proton dose calculations in adaptive proton therapy. *Phys Med Biol*. 2020;65:095002.

14. Thing R, Nilsson R, Andersson S, Berg M, Lund M. Evaluation of CBCT based dose calculation in the thorax and pelvis using two generic algorithms. *Physica Med*. 2022;103:157-165.
15. Yoo SK, Kim H, Choi BS, Park I, Kim JS. Generation and evaluation of synthetic computed tomography (CT) from cone-beam CT (CBCT) by incorporating feature-driven loss into intensity-based loss functions in deep convolutional neural network. *Cancers*. 2022;14:4534.
16. Liu J, Yan H, Cheng H, et al. CBCT-based synthetic CT generation using generative adversarial networks with disentangled representation. *Quant Imaging Med Surg*. 2021;11:4820.
17. Dhar T, Dey N, Borra S, Sherratt RS. Challenges of deep learning in medical image analysis—improving explainability and trust. *IEEE Trans Technol Soc*. 2023;4:68-75.
18. Emami H, Dong M, Nejad-Davarani SP, Glide-Hurst CK. Generating synthetic CTs from magnetic resonance images using generative adversarial networks. *Med Phys*. 2018;45:3627-3636.
19. Wu B, Xu C, Dai X, et al. Visual transformers: token-based image representation and processing for computer vision. *arXiv preprint arXiv:2006.03677*. 2020.
20. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*. 2020.
21. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Advances in Neural Information Processing Systems. Vol 30. 2017.
22. Tang Y, Yang D, Li W, et al. Self-supervised pre-training of swin transformers for 3D medical image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:20730-20740.
23. Xiao T, Singh M, Mintun E, Darrell T, Dollár P, Girshick R. Early convolutions help transformers see better. *Adv Neural Inf Process Syst*. 2021;34:30392-30400.
24. Dai Z, Liu H, Le QV, Tan M. CoAtNet: marrying convolution and attention for all data sizes. *Adv Neural Inf Process Syst*. 2021;34:3965-3977.
25. Wu H, Xiao B, Codella N, et al. CvT: introducing convolutions to vision transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:22-31.
26. Johnson H, Harris G, Williams K, et al. BRAINSFit: mutual information rigid registrations of whole-brain 3D images, using the insight toolkit. *Insight J*. 2007;57:1-10.
27. Fedorov A, Beichel R, Kalpathy-Cramer J, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. *Magn Reson Imaging*. 2012;30:1323-1341.
28. He K, Gan C, Li Z, et al. Transformers in medical image analysis: a review. *Intelligent Medicine*. 2022;3:59-78.
29. Liang J, Cao J, Sun G, Zhang K, Van Gool L, Timofte R. SwinIR: image Restoration Using Swin Transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1833-1844. 2021.
30. Schneider U, Pedroni E, Lomax A. The calibration of CT Hounsfield units for radiotherapy treatment planning. *Phys Med Biol*. 1996;41:111-124.

How to cite this article: Viar-Hernandez D, Molina-Maza JM, Vera-Sánchez JA, et al. Enhancing adaptive proton therapy through CBCT images: Synthetic head and neck CT generation based on 3D vision transformers. *Med Phys*. 2024;51:4922–4935. https://doi.org/10.1002/mp.17057