scientific reports

Check for updates

OPEN

CBCT-to-CT synthesis using a hybrid U-Net diffusion model based on transformers and information bottleneck theory

Can Hu[1], Ning Cao[1], Xiuhan Li[1,4,5], Yang He[1] & Han Zhou[2,3]✉

Cone-beam computed tomography (CBCT) scans are widely used for real time monitoring and patient positioning corrections in image-guided radiation therapy (IGRT), enhancing the precision of radiation treatment. However, compared to high-quality computed tomography (CT) images, CBCT images suffer from severe artifacts and noise, which significantly hinder their application in IGRT. Therefore, synthesizing CBCT images into CT-like quality has become a critical necessity. In this study, we propose a hybrid U-Net diffusion model (HUDiff) based on Vision Transformer (ViT) and the information bottleneck theory to improve CBCT image quality. First, to address the limitations of the original U-Net in diffusion models, which primarily retain and transfer only local feature information, we introduce a ViT-based U-Net framework. By leveraging the self-attention mechanism, our model automatically focuses on different regions of the image during generation, aiming to better capture global features. Second, we incorporate a variational information bottleneck module at the base of the U-Net. This module filters out redundant and irrelevant information while compressing essential input data, thereby enabling more efficient summarization and better feature extraction. Finally, a dynamic modulation factor is introduced to balance the contributions of the main network and skip connections, optimizing the reverse denoising process in the diffusion model. We conducted extensive experiments on private Brain and Head & Neck datasets. The results, evaluated from multiple perspectives, demonstrate that our model outperforms state-of-the-art methods, validating its clinical applicability and robustness. In future clinical practice, our model has the potential to assist clinicians in formulating more precise radiation therapy plans.

Precision radiotherapy seeks to optimize therapeutic radiation delivery by intensifying tumor exposure while sparing adjacent critical structures. Image-guided radiation therapy (IGRT) is pivotal in achieving this precision, enabling real-time monitoring of tumor and normal organ changes before and during treatment[1,2]. Computed tomography (CT), due to its superior imaging quality, has become a key diagnostic and IGRT tool in the medical field. However, relying solely on pre-treatment CT images can overlook the fact that patient anatomy may change during treatment, potentially leading to inaccurate radiation dose delivery to critical organs, reduced tumor control effectiveness, and increased risk of excessive radiation and critical organ doses[3–5]. Cone-beam computed tomography (CBCT) serves as an efficient daily or weekly verification tool, offering rapid acquisition and reduced exposure while facilitating precise patient positioning.

Despite its advantages, CBCT images often suffer from significant artifacts, lower image quality, and resolution, as well as inaccurate Hounsfield units (HU), which complicate the accurate identification of the boundary between normal organs and tumors. This hinders the quantitative application of CBCT and adaptive radiotherapy (ART) based on CBCT. Therefore, enhancing CBCT image quality is crucial. Enhanced CBCT imaging enables precise delineation of tumor-tissue interfaces, dynamically modify treatment plans during treatment, improve the accuracy of tumor radiotherapy, and better protect critical organs.

[1]School of Computer and Software, Hohai University, Nanjing 211100, China. [2]Department of Radiation Oncology, The Fourth Affiliated Hospital of Nanjing Medical University, Nanjing 210013, China. [3]School of Electronic Science and Engineering, Nanjing University, Nanjing 210046, China. [4]Jiangsu Province Engineering Research Center of Smart Wearable and Rehabilitation Devices, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China. [5]Engineering Research Center of Intelligent Theranostics Technology and Instruments, Ministry of Education, School of Biomedical Engineering and Informatics, Nanjing Medical University, Nanjing 211166, China. ✉email: hanzhou26@njmu.edu.cn

Current research on ART based on CBCT primarily focuses on two areas: artifact correction to improve CBCT image quality[6–21], and the generation of synthetic CT (sCT) images with quality comparable to CT[22–43]. Artifact correction can be broadly categorized into physical modeling-based and traditional algorithm-based methods, with deep learning methods playing a significant role in generative approaches. Physical modeling-based methods, for example, use physical modeling to simulate the interaction between scattered light and grating, thereby quantifying the grating's effect on removing scattered light[15]. Traditional algorithm-based methods include iterative image reconstruction algorithms that use sparse projection data for image reconstruction[16], optimization of Monte Carlo methods using scatter correction to improve image quality, and carbon fiber cross-type scatter-reducing gratings to reduce scattered radiation[17]. However, these methods struggle with large datasets and complex nonlinear relationships, leading to suboptimal results in tasks such as CBCT noise suppression and artifact removal. Consequently, researchers are exploring deep learning methods. Currently, deep learning techniques for generating sCT primarily involve generative adversarial networks (GANs), which consist of a generator and a discriminator, mapping CBCT to CT image domains to produce high-quality sCT images. Despite their potential, GANs, due to their adversarial training strategy, can suffer from instability, early convergence, and mode collapse, limiting the quality and diversity of synthetic images. Therefore, variants such as CycleGAN[28] and AttentionGAN[29] are increasingly used in CBCT-to-CT image conversion tasks, enhancing the accuracy and anatomical consistency of generated images through cycle consistency loss and attention mechanisms.

Recent years have witnessed the significant success of diffusion models as an alternative to GANs in the field of computer vision, particularly in the domain of image generation[44,45]. Unlike GANs, which do not depend on adversarial training techniques, diffusion models enhance the stability of training and produce more realistic output images with improved quality and greater semantic diversity. This approach is less time-consuming for fine-tuning the network and adjusting training hyperparameters. Drawing inspiration from non-equilibrium thermodynamics in physics, diffusion models represent a generative method that introduces random noise gradually to destroy images, defining a Markov diffusion step chain, and subsequently learns the inverse diffusion process to gradually restore generated samples from noise through denoising steps. Typically, the denoising network is constructed as a U-Net architecture, which is based on convolutional modules. The encoder of the U-Net extracts multi-scale features through downsampling, while the decoder reconstructs the image through upsampling, yielding denoising results. Despite its reliance on convolutional operations for feature extraction, the model's receptive field is limited, primarily capturing local information and lacking an effective global feature extraction mechanism. This limitation hampers the understanding of the image's overall semantics and long-range dependencies, leading to inconsistent and less precise detail reproduction in the generated results. Consequently, this affects visual quality and realism, particularly in the context of high-quality medical image generation or complex synthetic tasks.

In response to these deficiencies, we present a novel Hybrid U-Net Diffusion Model (HUDiff) that integrates Vision Transformer (ViT) architecture with information bottleneck principles to synthesize superior-quality sCT reconstructions from CBCT data. Our approach is structured into two phases: the first involves a forward noise addition process, where Gaussian noise progressively perturbs the initial CT data distribution over T iterations; followed by an inverse restoration phase utilizing a hybrid U-Net architecture to estimate perturbation at each iteration, with CBCT serving as conditional input to the U-Net to facilitate the generation of clear CT images. Specifically, our contributions include:

1. We propose a ViT-based U-Net framework to address the limitations of the original U-Net, which primarily retains and transfers local feature information. By leveraging the characteristics of self-attention mechanisms, our model adaptively prioritizes diverse spatial areas throughout synthesis, enhancing comprehensive feature interpretation.
2. We introduce a variational information bottleneck (VIB) module at the base of the U-Net, which filters out redundant and irrelevant information while compressing essential input data, thereby effectively summarizing the input and capturing key features more accurately.
3. We incorporate a dynamic modulation factor to balance the contributions of the main network and skip connections, optimizing the reverse denoising process in conjunction with the diffusion model.
4. We conduct a comprehensive analysis and comparison of the sCT images synthesized through HUDiff against contemporary approaches using diverse evaluation metrics, demonstrating the superiority of the HUDiff approach.

Related work

Device-level CBCT enhancement approaches

System-level CBCT enhancement approaches concentrate on optimizing scanner hardware configurations. This includes improvements in detector design, adjustments to the light source, the addition of filters, or the introduction of auxiliary correction devices, all aimed at reducing the impact of scattered radiation and artifacts on image quality, thereby enhancing accuracy and clarity. Zhu et al.[19] introduced an additional scattering detector array to directly measure scattered radiation, effectively mitigating the influence of scattering artifacts on CBCT images. Stankovic et al.[20] proposed an iterative model based on beam-blocking methods that concurrently reduces scattering from various grids, resulting in diminished artifact generation and improved CBCT image quality. Although hardware-based methods can mitigate scattering artifacts to some extent, challenges such as the limitations of physical equipment, the complexity of hardware implementation, and high costs remain significant obstacles.

Conventional computational approaches for CBCT enhancement

Traditional algorithm-based CBCT image correction methods involve applying techniques such as filtering, interpolation, regression, or model reconstruction in the projection or image domain to reduce artifacts, scattering noise, or other distortions, thereby enhancing the quality of CBCT images. Sidky et al.[16] developed an iterative image reconstruction algorithm based on sparse angles and limited views, which utilizes sparse projection data to reconstruct images and improve quality. Xu et al.[17] proposed a practical CBCT scattering correction method based on optimized Monte Carlo simulations to enhance CBCT image quality. Usui et al.[18] conducted Monte Carlo simulation studies using carbon fiber cross-type anti-scatter grids to reduce the impact of scattered radiation on CBCT image quality. Wang et al.[21] employed a prior-information-guided sequential random forest algorithm to develop an automated segmentation method for CBCT images, effectively improving segmentation accuracy and indirectly enhancing image quality. However, these algorithms struggle to optimize from input to output directly and typically require longer computation times, limiting the overall performance of image generation.

GAN-based image synthesis methods

Advanced neural architecture composed of dual complementary units a synthesis module and an evaluation module forms the foundation of GAN. The synthesis module creates realistic outputs from input sequences, while the evaluation component validates authenticity through comparative assessment. This adversarial optimization process progressively enhances output fidelity. This framework excels in cross-modality transformation tasks, particularly in medical imaging domain transitions. Liang et al.[30] pioneered a cycle-consistent synthesis approach for CT reconstruction from CBCT acquisitions in unsupervised environments. Empirical evaluations demonstrate this methodology's superior efficacy compared to contemporary unsupervised solutions. Building upon these advances, Sun et al.[31] established a protocol for sCT synthesis utilizing enhanced cyclical adversarial mechanisms. Their architecture incorporates reinforced U-Net pathways with integrated attention mechanisms and dimensional restoration functions to optimize reconstruction quality. Deng et al.[32] established a reconstruction protocol utilizing an enhanced ResPath architecture, incorporating residual circuits for high-fidelity sCT synthesis from CBCT acquisitions. Zhang et al.[33] advanced volumetric reconstruction frameworks by integrating physical principles with neural architectures to enhance resilience against data uncertainty. Li et al.[34] developed a spatially-aware synthesis framework (RegGAN) that harmonizes geometric alignments through U-Net integrations, addressing multi-platform reconstruction variations and HU inconsistencies. Semul et al.[35] implemented comprehensive skip-connection mechanisms for architectural optimization. Liu et al.[36] introduced a sequence-oriented synthesis methodology with adaptive temporal fusion, augmenting inter-modality consistency. Hu et al.[37] incorporated self-attention mechanisms into cross-modality synthesis to enhance spatial feature comprehension. Sun et al.[35] conceived a tri-modal synthesis architecture (TGAN) unifying CBCT, MRI, and CT modalities. Zhang et al.[39] established a comprehensive unsupervised transformation framework addressing stylistic elements in latent representations.

Diffusion model-based image synthesis methods

Diffusion models, a novel approach to generative methods, have demonstrated superior performance over GAN in image synthesis, denoising, and super-resolution tasks[46,47]. In the field of CBCT-to-CT synthesis, obtaining paired patient data is highly challenging and costly. Unsupervised or self-supervised diffusion models have gained significant attention as they can reduce dependence on paired data, making training data more accessible and facilitating broader research applications. The diffusion model operates in two phases: first incorporating Gaussian noise progressively into the CT image, followed by a reverse phase that estimates and eliminates the introduced noise, facilitating CBCT to CT image conversion. Li et al.[48] proposed a frequency-domain guided diffusion model FGDM, which utilizes a frequency-domain filter to guide the diffusion process and effectively preserves the structural information in the image. The model is trained using only data from the target domain without source domain data, realizing direct image conversion from source to target domain. The experimental results show that FGDM outperforms existing GAN, VAE and diffusion models in several medical image conversion tasks with significant advantages. Fu et al.[40] established EGDiff, a framework that leverages an energy conduction function to preserve domain-independent characteristics throughout the noise reduction sequence, aimed at optimizing sCT image fidelity. Peng et al.[41] implemented a probabilistic diffusion architecture for conditional denoising, which achieved exceptional results in dose calculation and visual quality metrics when utilizing CBCT images as conditions. Chen et al.[42] developed an adaptive diffusion-based architecture called L-DM, capable of accurately synthesizing the HU values of corresponding sCT images and effectively removing artifacts, advancing the clinical treatment of lung cancer radiotherapy. Zhang et al.[43] designed a fusion model that integrates attentive frequency optimization and a dual-branch feature module to enhance the high-frequency details of sCT and preserve details effectively, further advancing CBCT synthesis.

ViT and IB-MLP module in medical images

ViT has demonstrated significant advantages in capturing global features of images through its self-attention mechanism, and has been widely applied in various medical image tasks[49,50]. Unlike traditional CNN, which primarily rely on the transmission and retention of local information, ViT is capable of capturing long-range dependencies within an image, thereby enhancing its understanding of global features. Through self-attention, ViT dynamically focuses on different regions of an image, not only improving its clarity and realism but also compensating for the limitations of convolution operations. Manzari et al.[51] introduced a robust and efficient hybrid model that integrates the locality of convolutions with the global connectivity of ViT, designing an efficient attention mechanism and learning smoother decision boundaries. This approach achieved superior performance in medical image classification tasks. Dalmaz et al.[52] proposed a novel generative adversarial model, ResViT,

which combines the contextual sensitivity of ViT, the precision of convolutional operations, and the realism of adversarial learning. Extensive experiments on multi-contrast MRI sequence synthesis and MRI-to-CT image generation demonstrated that ResViT outperforms existing methods both qualitatively and quantitatively. He et al.[53] introduced an efficient hierarchical hybrid ViT model for medical image segmentation, which effectively utilizes limited medical data. The experimental results showed that this model outperforms existing methods in three 2D and two 3D medical image segmentation tasks, while maintaining high computational efficiency in terms of model parameters, FLOPs, and inference time.

Information bottleneck theory improves feature extraction by compressing input data and retaining only the task-relevant key information[54]. The IB-MLP module, based on this theory, has been widely integrated into various models to enhance the quality of medical images. The IB-MLP module filters redundant information and compresses data, removing unnecessary anatomical features. This results in improved tissue contrast, fewer artifacts, and more accurate anatomical structure delineation, thereby increasing the precision of feature extraction. Li et al.[55] innovatively incorporated Transformer and IB modules into the U-Net model, using the IB module to compress redundant features and mitigate the risk of overfitting. Ablation studies conducted on two public datasets, along with comparisons to state-of-the-art models, demonstrated the advantages of their approach. Furthermore, Li et al.[56] proposed a hybrid model based on diffusion models and the IB-MLP module. This model combines the denoising and global feature-capturing abilities of diffusion models with the information compression capabilities of the IB-MLP module. Located at the bottom of the model, the IB-MLP module uses information bottleneck theory to compress learned features, retaining the most relevant features for the task while discarding irrelevant ones, thus improving the model's generalization ability. Experimental results on three public datasets, compared with leading models, show that this approach offers significant advantages in medical image segmentation tasks.

Based on the above study,we propose the HUDiff model,a hybrid U-Net diffusion model for synthesizing CT images from CBCT images. We validate the effectiveness of HUDiff on a private dataset, and the experimental results show that the method is able to accomplish the task of image synthesis from CBCT to CT with high quality, and exhibits good performance in medical imaging applications.

Method

In this section, we first introduce the framework of the conditional diffusion model. Next, we present the overall architecture of the hybrid U-Net denoising network within the conditional diffusion model. Finally, we discuss the information bottleneck theory and dynamic modulation factor utilized in the hybrid U-Net.

Conditional diffusion model

As shown in Fig. 1, x_0 is the CT image and y is the CBCT image. We first add noise to x_0, continuously in the forward process, and set the step size to T, we can obtain $x_1, x_2, \ldots x_t$, where the added noise obeys the Gaussian distribution $\mathcal{N}(0, I)$, So x_t, the distribution can be expressed as follows:

$$q(x_t|x_0) = \mathcal{N}\left(x_t; \sqrt{\overline{\alpha}_t}x_0, (1-\overline{\alpha}_t)I\right) \tag{1}$$

where $\overline{\alpha}_t := \prod_{i=1}^{t} \alpha_n$ and $\alpha_t := 1 - \beta_t$ with β_t being the variance of the forward process and I representing the identity matrix. Ultimately, x_t can be expressed as a function of x_0 as follows:

$$x_t = \sqrt{\overline{\alpha}_t}x_0 + \sqrt{1-\overline{\alpha}_t}\varepsilon, \quad with\ \varepsilon \sim \mathcal{N}(0, I) \tag{2}$$

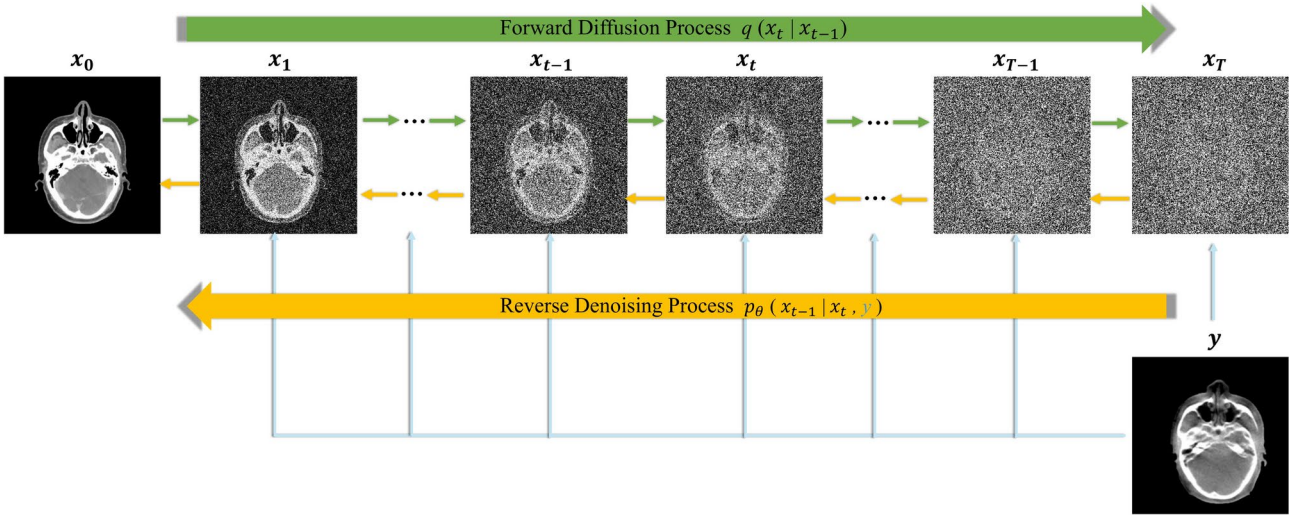


Fig. 1. General framework for conditional diffusion modeling.

Secondly, we denoise x_t. in the reverse process, but $q(x_{t-1}|x_t)$ Unknown, so we use the Bayesian formula to solve:

$$q(x_{t-1}|x_t, x_0) = q(x_t|x_{t-1}, x_0)\frac{q(x_{t-1}|x_0)}{q(x_t|x_0)} \tag{3}$$

Based on Eqs. (1) and (3), the distribution $q(x_{t-1}|x_t)$ can be expressed as:

$$q(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \overline{\mu}(x_t, t), \overline{\beta}I\right) \tag{4}$$

where

$$\overline{\mu} = \frac{\sqrt{\overline{\alpha}_{t-1}}\beta_t}{1-\overline{\alpha}_t}x_0 + \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t \tag{5}$$

and

$$\overline{\beta} = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t \tag{6}$$

From Eqs. (2), (5) can be rewritten as:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t.}}\varepsilon_\theta(x_t, t)\right) \tag{7}$$

Unlike other diffusion models, the conditional diffusion model used in this paper uses y_0 to guide x_0 for denoising in the inverse process, so x_t can be used to estimate x_{t-1}:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\varepsilon_\theta(x_t, y_0, t)\right) + \sigma_t z \tag{8}$$

Since the backward denoising process of the conditional diffusion model uses the U-Net architecture, the goal of this paper is to train the neural network parameter ε_θ to estimate the noise at each step, and the loss function of the network is defined as follows:

$$\mathscr{L}(\theta) = \mathbb{E}x_t, y_0, t[|x_{t-1} - y_0|2^2] + \lambda_t \mathbb{E}x_t, y_0, t\left[|\nabla_{x_{t-1}}\varepsilon_\theta(x_t, y_0, t)|_2^2\right] \tag{9}$$

where $\mathbb{E}$ denotes the expectation of the training data. The first term in the function is denoted to make x_{t-1} as similar as possible to y_0 and in the second term ε_θ is the gradient of x_{t-1} which is used to smooth out the noise estimate. λ_t is the weighting factor. Minimization of $\mathscr{L}(\theta)$ is done by backpropagation with gradient descent, aiming at generating a high quality y_0 image from a given x_0 image.

Hybrid U-Net denoising network architecture

The denoising network of the original diffusion model consists of an encoder and a decoder, both equipped with convolutional modules. In this structure, the encoder acts as the downsampling path, encoding the input image into features of different resolutions, while the decoder serves as the upsampling path, assembling the final denoised image from these features. Despite the U-Net's good performance, its structure has inherent limitations. The network primarily relies on convolutional operations, which have a relatively small receptive field and can only capture information from local areas. The lack of an effective global feature extraction mechanism makes the U-Net struggle to understand the overall semantic information and long-range dependencies in the image, leading to poor performance in image generation tasks, with generated images often lacking global consistency and detail realism. To address this issue, this study constructed a U-Net framework based on ViT, as shown in Fig. 2a, where the U-Net structure is used to extract and retain important anatomical features and detail information, and dynamic modulation factors are introduced to balance the contributions of the backbone network and skip connections. The self-attention mechanism of ViT is then utilized to automatically focus on different positions of the image during image generation, better understanding the global features of the anatomical image, making the generated images clearer and more realistic. Finally, a VIB module is introduced, as shown in Fig. 2c, to filter out redundant and irrelevant information and compress the key input data, aiming to effectively summarize the input and better capture key features. In the network architecture design, the U-Net adopts a symmetric encoder-decoder structure. Its encoding path includes four consecutive feature extraction layers, each consisting of a convolutional operation and a downsampling module. To effectively fuse features of different scales, the network design a skip connection mechanism, directly transferring feature maps extracted from each layer of the encoding path to the corresponding level of the decoding path. Specifically, the input image is first transformed into a feature tensor of size (w_0, h_0, f_0) through a preprocessing module. Subsequently, the tensor is processed through the downsampling modules in sequence on the encoding path, with the spatial dimensions of the feature maps halved at each module while the channel dimension f_0 is expanded to double. The ViT module, as shown in Fig. 2b, is mainly composed of multiple transformer encoder blocks, with an input feature dimension of $(w_0/16, h_0/16, 8f_0)$. The processing flow is as follows: the input feature map is first

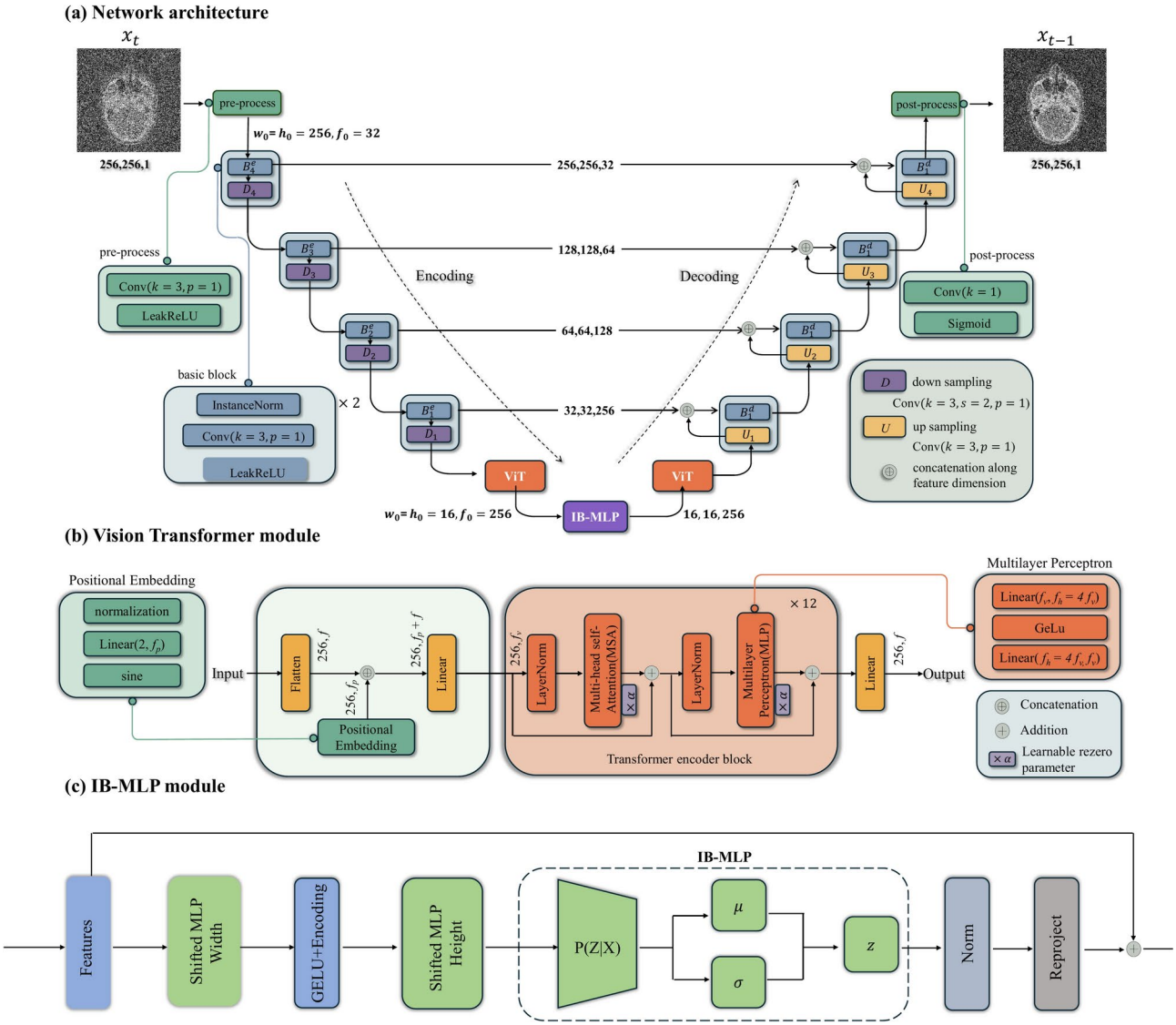


Fig. 2. Hybrid U-Net denoising network architecture.

rearranged into a token sequence in the spatial dimension, with a sequence length of $w \times h$, where each token is represented as a vector of length *f*. Subsequently, the token sequence is fused with a two-dimensional Fourier position encoding of dimension f_p, and then projected linearly to dimension f_v as the input to the encoder. To optimize the model training process, we adopted the Rezero regularization strategy and introduced learnable scale parameters α to dynamically regulate the contribution weights of the non-linear transformation branch in the residual connection, which significantly improved the convergence performance of the model.

IB-MLP module

The information bottleneck theory provides a theoretical basis for the feature extraction process in deep learning. The theory states that an effective learning system should be able to compress the input data and retain only the key information related to the task. By establishing the minimum sufficient statistic between the input and the target, the model can filter redundant features, thus enhancing its generalization performance. This learning paradigm based on information compression not only simplifies the feature representation, but also provides a new way of thinking to improve the model robustness. Its theory can be expressed as follows:

$$maxL_{IB}(Z) = I(Z;Y) - \varepsilon I(X;Z) \tag{10}$$

In this flalign, the system utilizes three stochastic variables: *X* denotes the input data, *Y* represents the output data, and *Z* characterizes the latent representation. The optimization objective involves tuning the parameter ε of the learning algorithm to achieve two complementary goals: maximizing $I(Z; Y)$—the mutual information between the derived representation *Z* and output *Y*, while simultaneously minimizing $I(Z; X)$—the corresponding information metric between *Z* and the input variable *X*.

The variational information bottleneck is modeled by a deep neural network for probability distributions and combines reparameterization techniques with Monte Carlo sampling methods to achieve unbiased gradient estimation. This design breaks through the limitations of traditional methods on discrete or Gaussian distributions, enabling them to effectively handle high-dimensional continuous data, and realizing the preservation of key information and compression of redundant features. In the concrete implementation, we integrate the VIB module into the bottom layer of U-Net architecture. This design is based on the following considerations: firstly, the bottom layer is the final stage of feature learning, at this time the compression of the fully learned features is more meaningful; secondly, due to the information loss in the upper layer of the network, if the feature compression in this position may affect the convergence performance of the model. the loss function of the IB module can be expressed as follows:

$$Loss_{IB-MLP} = maxI(Z;Y) - \varepsilon I(X;Z) \tag{11}$$

Here, X encodes the source features from the initial feature space f, while Z comprises a subset of X containing elements with maximal relevance to the target prediction Y. The coefficient ε spans [0,1], regulating the extent of dimensionality reduction. The quantities $I(Z; Y)$ and $I(X; Z)$ serve as information-theoretic metrics to evaluate cross-variable dependencies. This formulation aims to optimize a dual objective: enhancing $I(Z; Y)$ to preserve essential predictive elements while constraining $I(X; Z)$ to achieve compact representations and mitigate potential overfitting. The mathematical formalism of these information-theoretic components takes the following form:

$$I(Z;Y) = \int dydzp(y,z)log\frac{p(y|z)}{p(y)} \tag{12}$$

$$I(X;Z) = \int dzdxp(x,z)log\frac{p(z|x)}{p(z)} \tag{13}$$

Due to the high-dimensional distribution of image data, calculating $p(y|z)$ poses significant challenges. By leveraging the property that the Kullback-Leibler divergence (KLD) between two distributions is always non-negative, we learn the distribution $q(y|z)$ to approximate $p(y|z)$. The lower bound of $I(Z; Y)$ can be expressed as:

$$I(Z;Y) \geq \int dxdydzp(x)p(y|x)p(z|x)logq(y|z) \tag{14}$$

For $I(X; Z)$, we select a standard Gaussian distribution $r(z)$ to approximate $p(z)$. The upper bound of $I(X; Z)$ can be represented as:

$$I(X;Z) \leq \int dxdzp(x)p(z|x)log\frac{p(z|x)}{r(z)} \tag{15}$$

Thus, we can express $I(Z;Y) - \varepsilon I(X;Z)$ as follows:

$$\begin{aligned} I(Z;Y) - \varepsilon I(X;Z) \geq & \int dxdydzp(x)p(y|x)p(z|x)\log q(y|z) \\ & - \varepsilon \int dxdzp(x)p(z|x)\log\frac{p(z|x)}{r(z)} = L \end{aligned} \tag{16}$$

We substitute the empirical distribution $p(x,y) = \frac{1}{N}\sum_{n=1}^{N}\delta_{x_n}\cdot\delta_{y_n}$ for $p(x, y)$. Consequently, L and the loss function $Loss_{IB-MLP}$ can be reformulated as:

$$L \approx \frac{1}{N}\sum_{n=1}^{N}\left[\int dzp(z|x_n)\log q(y_n|z) - \varepsilon p(z|x_n)\log\frac{p(z|x_n)}{r(z)}\right] \tag{17}$$

Finally, we have:

$$Loss_{IB-MLP} = \frac{1}{N}\sum_{n=1}^{N}\left[\varepsilon KL(p_\phi(z|x_n) \parallel r(z)) - \mathbb{E}_{z\sim p_\phi(z|x_n)}[\log q(y_n|z)]\right] \tag{18}$$

Here, E denotes the expected value, and $p_\phi(z|x_n)$ represents a variational Gaussian distribution encoder that learns the mean μ and variance σ through the network.

Dynamic modulation factor

According to the paper by Si et al.[57], the primary function of the U-Net backbone network is to filter out high-frequency noise, thereby ensuring the fidelity and detail of the images. The skip connections forward features directly from the earlier layers of the encoder blocks to the decoder, which primarily consist of high-frequency signals. While the presence of these high-frequency features may accelerate the convergence of noise prediction,

it can also weaken the inherent denoising capability of the backbone network. To enhance the denoising ability of U-Net, we introduce two specialized dynamic modulation factors aimed at balancing the contributions of features from the U-Net backbone and the skip connections. The average feature map along the channel dimension is represented as:

$$\overline{x}_l = \frac{1}{C}\sum_{i=1}^{C} x_{l,i} \tag{19}$$

Here, $x_{l,i}$ denotes the *i*-th channel of the backbone feature map x_l in the *l*-th block of the U-Net decoder, and *C* represents the total number of channels in x_l. The backbone factor α_l amplifies the backbone feature map x_l in a manner consistent with its structural features. The backbone factor is defined as:

$$\alpha_l = (b_l - 1) \cdot \frac{\overline{x}_l - Min(\overline{x}_l)}{Max(\overline{x}_l) - Min(\overline{x}_l)} + 1 \tag{20}$$

where b_l >1. To achieve the most effective denoising results, we restrict the scaling operation to half of the channels in x_l:

$$\boldsymbol{x}'_{l,i} = \begin{cases} \boldsymbol{x}_{l,i} \odot \boldsymbol{\alpha}_l & \text{if } i < C/2 \\ \boldsymbol{x}_{l,i} & \text{otherwise} \end{cases} \tag{21}$$

To further mitigate the issue of excessive smoothing of textures due to denoising, we employ spectral modulation in the Fourier domain to selectively reduce the low-frequency components of the skip connection features. This process is as follows:

$$\begin{aligned} \mathscr{F}(h_{l,i}) &= \mathrm{FFT}(h_{l,i}) \\ \mathscr{F}'(h_{l,i}) &= \mathscr{F}(h_{l,i}) \odot \beta_{l,i} \\ h'_{l,i} &= \mathrm{IFFT}(\mathscr{F}'(h_{l,i})) \end{aligned} \tag{22}$$

In these flaligns, $h_{l,i}$ denotes the *i*-th channel of the skip feature map in the *l*-th block of the U-Net decoder, while FFT and IFFT represent the Fourier Transform and Inverse Fourier Transform, respectively. The term $\beta_{l,i}$ serves as the Fourier mask, with the frequency-dependent scaling factor s_l defined as:

$$\beta_{l,i}(r) = \begin{cases} s_l & \text{if } r < r_{\text{thresh}}, \\ 1 & \text{otherwise.} \end{cases} \tag{23}$$

Here, *r* is the radius, and r_{thresh} is the threshold frequency, which is set to 1 in our experiments.

Statement

We confirm that all methods were performed in accordance with the relevant guidelines and regulations, and informed consent for patients was waived by the Research Ethics Committee of the Nanjing Medical University.

Experiment

Dataset

This study collected CT and CBCT images from 105 patients with brain tumors and 108 patients with head and neck (H&N) cancers who underwent radiotherapy at the Fourth Affiliated Hospital of Nanjing Medical University, between October 1, 2021, and September 1, 2024. Specifically, we selected 9350 paired images for training from 85 brain tumor patients and randomly selected 1000 paired images for testing from 20 brain tumor patients. For the H&N group, we chose 8652 paired images for training from 84 patients and randomly selected 1200 paired images for testing from 24 patients. In this study, two imaging systems were used to acquire patient image data: a GE large-aperture CT device and an Elekta Exesse XVI system. The CT images were acquired with the following parameters: CT image voxel size of $0.625 \times 0.625 \times 1$ mm^2, size of 512×512, and slice thickness of 1.75 mm. CBCT scanning was performed with the following parameter settings: 360-degree rotation of the frame, 120 kVp, 10 mA, 10 ms, and F0S20 collimator. The acquired CBCT images had a voxel size of $1 \times 1 \times 1$ mm^2 and an axial dimension of 410×410. During the radiotherapy procedure, each patient first underwent a planned localization CT scan, followed by three CBCT image acquisitions during each weekly treatment period.

Data preprocessing

During clinical radiographic acquisition, extraneous medical equipment including treatment platforms, patient immobilization apparatus and protective gear—inadvertently appear in both CT and CBCT scans. The presence of these auxiliary components impacts both computational efficiency during model development and degrades output quality. Our methodology addresses this limitation through a two-phase mask-based purification protocol: Initially, medical professionals delineate regions of interest, followed by the application of customized filtering algorithms to extract non-anatomical elements. In the subsequent registration phase, inherent variations in equipment specifications and scanning protocols often result in spatial inconsistencies between corresponding

Methods	Brain				H&N			
	MAE	PSNR	SSIM	NCC	MAE	PSNR	SSIM	NCC
CycleGAN[25]	30.7425	28.1578	0.9525	0.9588	33.7715	26.1336	0.9474	0.9502
ADCycleGAN[58]	28.4446	28.9856	0.9602	0.9611	31.4552	27.3545	0.9558	0.9541
IViTCycleGAN[37]	26.2135	29.2355	0.9714	0.9634	30.4821	28.4815	0.9615	0.9601
DDPM[41]	25.3858	29.7879	0.9737	0.9702	28.4101	29.4448	0.9664	0.9689
EGDiff[40]	24.1379	30.4543	0.9788	0.9755	27.0179	30.0429	0.9711	0.9713
Ours	23.3114	31.3410	0.9862	0.9831	26.1114	30.8847	0.9802	0.9798

Table 1. Quantitative results of sCT synthesized by different models on Brain and H&N datasets.

Methods	Brain				H&N			
	MAE	PSNR	SSIM	NCC	MAE	PSNR	SSIM	NCC
CycleGAN[25]	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
ADCycleGAN[58]	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
IViTCycleGAN[37]	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
DDPM[41]	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
EGDiff[40]	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Ours	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001

Table 2. Indicator significance of different models versus our model on Brain and H&N datasets.

CT-CBCT pairs. Given the inherent geometric consistency of anatomical features, we implemented Advanced Normalization Tools (ANTS), an established registration framework, to resolve these alignment disparities.

Evaluation metrics and network training details

To quantitatively evaluate the fidelity between model-synthesized and reference CT volumes, we adopted a comprehensive suite of image quality metrics. The evaluation protocol encompasses Mean Absolute Error (MAE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Normalized Cross-Correlation[41] (NCC). In this study, all input images are preprocessed: the pixel values are normalized to the interval $[-1,1]$ and uniformly scaled to 256×256 resolution. The experiments involve two types of algorithm sets: CycleGAN-based methods (including the original CycleGAN, ADCycleGAN and IViT-CycleGAN) and diffusion models (DDPM, EGDiff and the algorithm proposed in this paper). For the first type of algorithms, we use the following training parameters: initial learning rate $r = 0.0002$, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for Adam optimizer, training rounds set to 200, and 2 samples processed per batch. The diffusion model was configured as follows: time step $T = 1000$, initial and termination noise variance $\beta_0 = 0.0004$ and $\beta_T = 0.02$, respectively. All algorithms were implemented based on Python 3.6 and TensorFlow 1.16 frameworks, and the training was done on a Linux server equipped with 8 NVIDIA Tesla V100 graphics cards.

Result

Comparison of different methods

Table 1 demonstrates the performance comparison of the proposed method with the existing state-of-the-art algorithms. In the Brain dataset, the proposed method achieves the optimal performance in all the indexes: the MAE is 23.3114, the PSNR reaches 31.3410, the SSIM obtains 0.9862, and the NCC reaches 0.9831. In order to validate the versatility of the algorithm, we apply the same parameter configurations to the H&N dataset, and we also obtain the excellent results: the MAE is 26.1114, the PSNR reaches 30.8847, the SSIM obtains 0.9802, and the NCC achieves 0.9798. These results are the same as those in Table 1. 26.1114, PSNR of 30.8847, SSIM of 0.9802, and NCC of 0.9798, which confirmed the leading edge of the proposed method in sCT image quality and demonstrated good cross-site adaptability. Further, we used t-test ($P < 0.05$) to assess the significance of the performance difference between the proposed method and other algorithms. As shown in Table 2, the statistical analyses on both datasets indicate that the proposed method has a significant advantage over the existing algorithms, which fully validates the advancement and generalizability of the method.

Ablation studies

In order to verify the effectiveness of each module, we conducted ablation experiments by gradually superimposing different functional modules using DDPM as the baseline network. The three core modules proposed in this paper include: Vision Transformer (ViT), Information Bottleneck MLP (IB-MLP) and Dynamic Modulation Factor (DMF). As shown in Table 3, the experimental results on both Brain and H&N datasets indicate that each module contributes positively to the model performance improvement. Specifically, the ViT module utilizes its built-in self-attention mechanism to adaptively capture the long-range dependencies of the image, which enhances the model's ability to understand the global semantic features, thus improving the clarity and realism of the generated images. The IB-MLP module, which is set in the bottom layer of U-Net, realizes the extraction

ViT	IB-MLP	DMF	Brain				H&N			
			MAE	PSNR	SSIM	NCC	MAE	PSNR	SSIM	NCC
		✓	25.3315	29.8541	0.9745	0.9715	28.0259	29.5849	0.9680	0.9701
	✓		25.2213	30.1225	0.9759	0.9733	27.8845	29.7462	0.9715	0.9715
✓			25.0131	30.3361	0.9772	0.9745	27.6124	29.8863	0.9729	0.9733
	✓	✓	24.5754	30.4551	0.9802	0.9787	27.3322	29.9910	0.9744	0.9751
✓		✓	24.0513	30.7715	0.9827	0.9801	27.0984	30.3615	0.9778	0.9769
✓	✓		23.8457	31.1251	0.9838	0.9811	26.8847	30.5847	0.9793	0.9781
✓	✓	✓	23.3114	31.3410	0.9862	0.9831	26.1114	30.8847	0.9802	0.9798

Table 3. Performance metrics of DDPM component ablation studies on Brain and H&N cohorts.

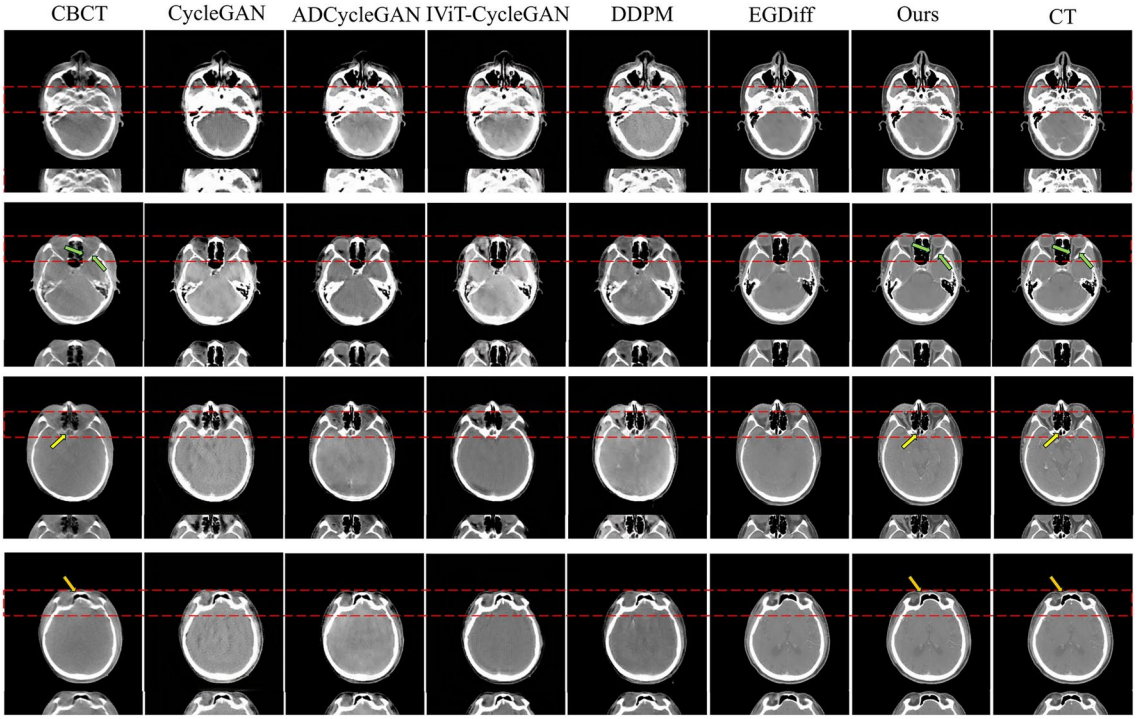


Fig. 3. Qualitative analysis of sCT reconstruction outcomes from the Brain dataset validation cohort.

of key features and the filtering of redundant information through the information compression mechanism to optimize the effectiveness of feature representation, and the DMF module dynamically adjusts the feature fusion weights of the backbone network and the hopping connection to promote the denoising effect of the diffusion model and guarantee the quality of the image generation. By comparing the four evaluation indexes of MAE, PSNR, SSIM and NCC, it is found that the ViT module contributes most significantly to the performance improvement, which is better than the IB-MLP and DMF modules.

Visual presentation and subjective evaluation

In addition to the quantitative evaluation metrics mentioned above, we further validated the performance of our method through multi-angle visualizations and by comparing the generated results from different models. Figure 3 compares the synthetic results of six algorithms CycleGAN, ADCycleGAN, IViT-CycleGAN, DDPM, EGDiff, and Ours on a Brain dataset, arranged from left to right. The anatomical structures chosen, from top to bottom, include the skull, neural tissue (green arrow), pituitary region (yellow arrow), and frontal sinus and orbit (orange arrow). From the magnified image portions, it is clear that the other comparative algorithms introduce significant noise and artifacts, and fail to accurately capture fine anatomical details, with blurred boundary contours. Specifically, in the first row of the Brain dataset, the skull region generated by CycleGAN and its variants, ADCycleGAN and IViT-CycleGAN, shows significant morphological differences and severe detail loss compared to real CT images, with prominent noise in the overall images. Although DDPM and its improved version EGDiff exhibit less detail loss compared to real CTs, they still contain some noise and artifacts that could affect clinical diagnosis. In the second row, for the neural tissue (green arrow), CycleGAN shows large morphological differences compared to real CTs, failing to generate neural tissue altogether.

ADCycleGAN and IViT-CycleGAN are able to generate partial tissue but suffer from detail loss, feature errors, and morphological distortions. DDPM and EGDiff show smaller morphological differences compared to real CTs, but the orbit region still exhibits slight detail loss and noise. In the third row, for the pituitary region (yellow arrow), CycleGAN and its variants produce a noticeably blurred pituitary gland and unclear orbital boundaries with significant detail loss. DDPM and EGDiff show minimal detail loss compared to real CTs, but the orbital region remains blurry, making clinical diagnosis challenging. In the fourth row, for the frontal sinus and orbit (orange arrow), CycleGAN and its variants exhibit blurred boundaries and incorrect structures compared to real CTs. DDPM and EGDiff show smaller morphological differences, but there are still minor distortions and noise that could interfere with targeted radiotherapy. Most of these structures are bony and stable, which are often used for registration during radiotherapy. Accurate generation of these structural details is crucial for registration outcomes and clinical treatment implementation. In contrast, our method produces images with fewer noise and artifacts, more accurate and realistic details, and clearer boundaries, closely resembling real CT results. We also compared the synthetic results of the six algorithms on the H&N dataset, as shown in Fig. 4. The anatomical structures, from top to bottom, include the nasopharynx (green arrow), mandible (yellow arrow), orbit (orange arrow), and cervical spine. From the magnified portions of the images, it is evident that the comparative algorithms again introduce significant noise and artifacts, with substantial loss of anatomical detail surrounding the lesions and blurred boundary contours. Specifically, in the first row of the H&N dataset, CycleGAN and its variants, ADCycleGAN and IViT-CycleGAN, show significant detail loss and boundary deformations in the nasopharynx, with considerable image distortion. DDPM produces less detail loss, but the contours are significantly distorted. EGDiff compensates for contour distortions but still suffers from detail loss, which may lead to clinical errors and treatment deviations. In the second row, in the mandible region, CycleGAN and its variants generate incorrect structures with large differences from real CTs, losing significant contour information and rendering the images too blurred to assist clinical diagnosis. DDPM and EGDiff show more similarity to real CTs in the mandible region, but still exhibit some detail loss and noise artifacts in other areas. In the third row, CycleGAN fails to generate the orbit region entirely, resulting in severe distortions. Although ADCycleGAN and IViT-CycleGAN generate the orbit region, the boundaries are blurred and severe detail loss and artifacts are present elsewhere. DDPM and EGDiff perform better in generating the orbit region compared to previous algorithms but still show minor boundary blurriness and artifacts. In the fourth row, CycleGAN shows substantial detail loss and blurriness in the cervical spine region. Although ADCycleGAN, IViT-CycleGAN, and DDPM perform better in this region, the overall images are still somewhat blurred, with regional detail loss and noise. EGDiff produces images closer to real CTs overall, but there are still minor differences in the cervical spine details. While this region may have a lesser impact on lesion detail generation, it is crucial for clinical registration, which directly affects radiotherapy positioning accuracy and treatment outcomes. In contrast, our method generates images with fewer regional artifacts, more accurate and realistic details, and clearer boundaries, offering imaging references that support adaptive radiotherapy to some extent.

Figure 5a and b present the results of the comparative difference analysis of the two datasets. The rainbow chromatograms are shown, and the degree of difference is indicated by the gradation from blue (minimum) to

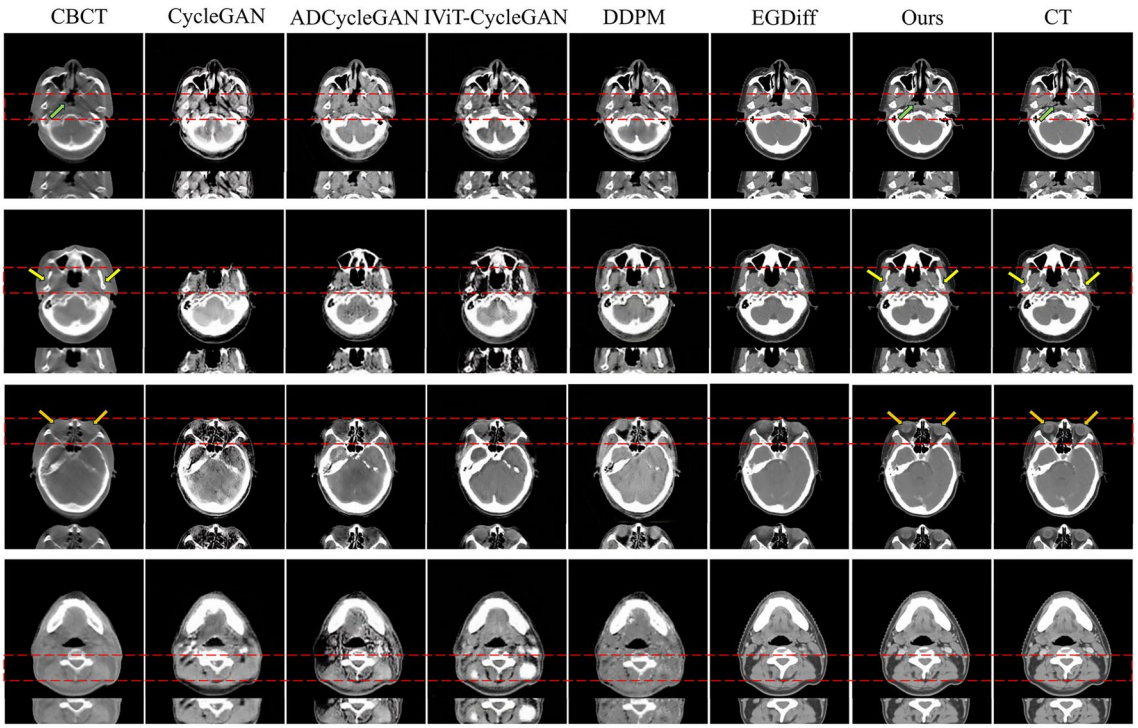
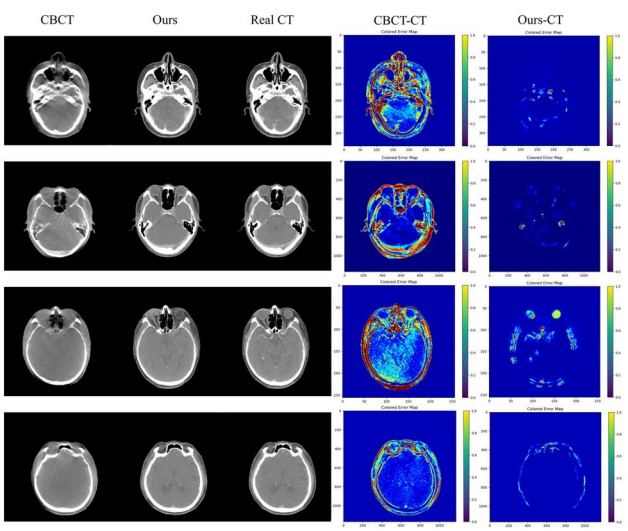
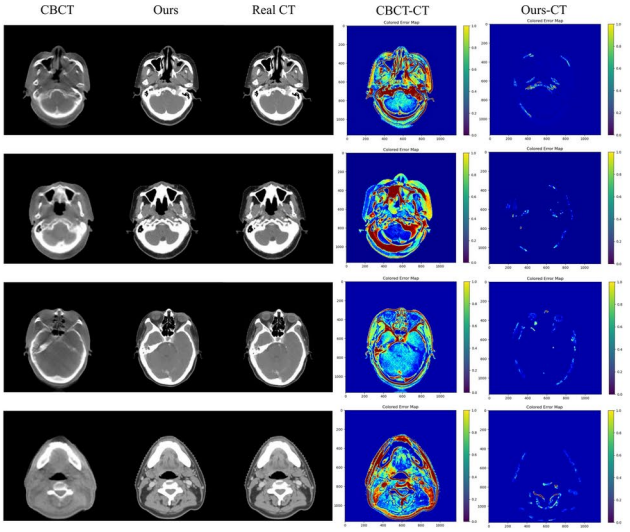


Fig. 4. Qualitative analysis of sCT reconstruction outcomes from the H&N dataset validation cohort.

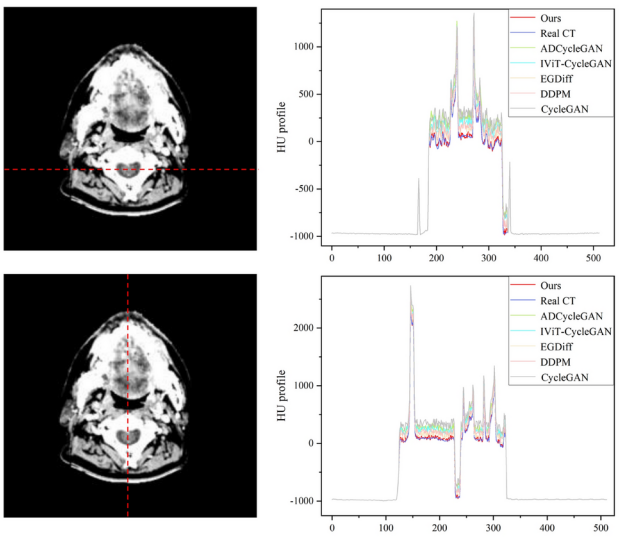


(a) Error plots of CBCT and CT and the method of this paper versus CT for the Brain dataset.

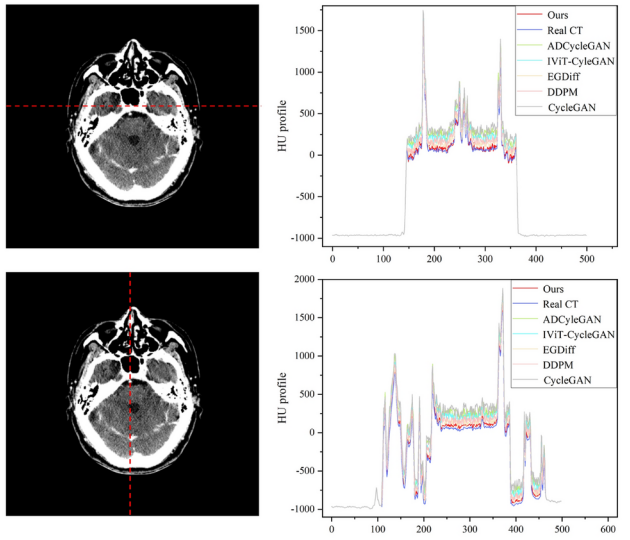


(b) Error plots of CBCT and CT and the method of this paper versus CT for the H&N dataset.

Fig. 5. Error plots comparison for Brain and H&N datasets.



(a) Statistical characterization of Hounsfield unit distributions across the Brain dataset validation cohort.



(b) Statistical characterization of Hounsfield unit distributions across the H&N dataset validation cohort.

Fig. 6. CT value distribution visualizations for the Brain and H&N dataset test sets.

red (maximum). The quantitative assessment demonstrates that our proposed approach achieves substantially reduced discrepancy from reference CT compared to conventional CBCT scans, indicating superior image quality approximating diagnostic CT standards.

In addition to the qualitative assessment of the details of tissue generation, this study also quantitatively analyzes the performance of each model through the distribution of CT values, which reflect the degree of absorption of X-rays by tissues and their relative density characteristics. We averaged the pixel values on the vertical and horizontal axes for each sample in the test set, and the resulting distributions are shown Fig. 6a and b The horizontal axis indicates the pixel position and the vertical axis indicates the average value, in which the purple curve represents the true CT distribution, the red curve is the result of the present method, and the other colors correspond to the rest of the comparison methods. The analysis shows that, although the distribution trends of all the methods are basically consistent with the real CT, the distribution curves of the present method are closer to the real CT than the other generally high results, reflecting the better realism.

In addition to the evaluation of objective indicators, we also conducted a subjective evaluation. The subjective evaluation was performed by two senior radiotherapists. The two observers analyzed and scored the mixed

images of CBCT, CT, and sCT at three time points (each three weeks apart) in randomized order, using the Likert scale, in terms of clarity of bony structures, detail of local organ segmentation, distortion of images, and overall image quality, without knowledge of the corresponding patients (with a score of 1 as the worst image quality and 5 as the best image quality). (1 being the worst image quality and 5 being the best image quality). The subjective quality scores of the three groups of images by the two reviewers are shown in Table 4. CT and sCT had higher scores than CBCT in terms of clarity of bone structure clarity, detail of local organ segmentation, image deformation distortion, and overall image quality, and the difference was statistically significant ($P < 0.001$). While the difference between CT and sCT in terms of scores was not statistically significant ($P>0.05$). This indicates that sCT is close to the reference image CT in terms of subjective clinical scoring and has a very high degree of authenticity.

Tumor target area outlining and dose distribution

CT imaging, based on the principle of X-ray, captures two-dimensional slice data. However, in clinical practice, three-dimensional reconstruction techniques are often required to observe tumor information from multiple dimensions in order to enhance diagnostic and therapeutic accuracy. To validate the image quality of the generated sCT, this study performed three-dimensional reconstruction and evaluated its spatial continuity. Figure 7 demonstrates the tumor target region contouring results for the Brain and H&N datasets across three planes: axial, sagittal, and coronal. The axial slices are directly output from the model, while the other two planes are reconstructed in three dimensions. Figure 7 shows tumor region contours manually drawn by experienced radiation oncologists on both synthetic CT and real CT images, followed by fusion of the two. Figure 7a illustrates the similarity of the target region contours for the Brain dataset, where the red solid line represents the real CT, and the yellow solid line represents the synthetic CT. It is evident that the similarity of the target regions is high, indicating that the synthetic CT image quality is highly similar to the real CT. Figure 7b demonstrates the target region contouring similarity for the H&N dataset, where the red solid line represents the real CT and the green solid line represents the synthetic CT. While the target region similarity remains high, it is slightly lower than that observed in the Brain dataset. This discrepancy is due to the presence of artifacts in the real CT images, commonly caused by metal or dental structures in the H&N region, which leads to increased noise. Furthermore, H&N scans often use artifact-reducing sequences and cover a larger scan range, resulting in less reconstruction information in the real CT images. The reconstruction results presented in Figs. 8 and 9 show that the synthetic CT not only preserves the original anatomical features but also maintains good continuity of tissue across all dimensions. Additionally, the dose-volume histogram (DVH) comparison results on the right side of Figs. 8 and 9, with solid and dashed lines representing the real CT and the proposed method's distributions, respectively, demonstrate a high degree of agreement between the two. This further validates the potential application of this method in clinical radiotherapy planning.

Dose calculation

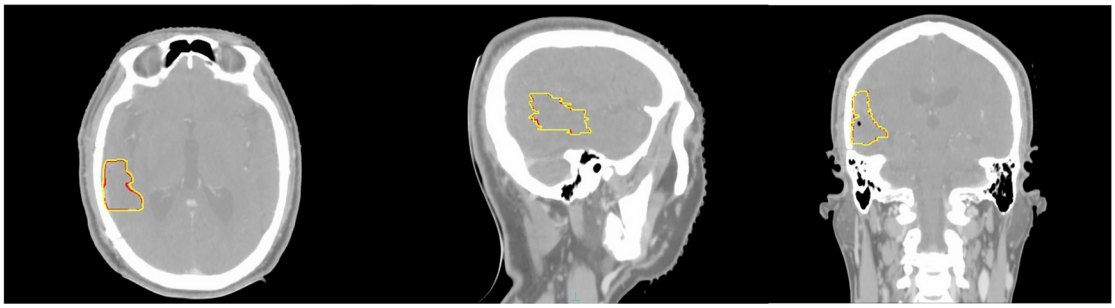
Dose calculation, as the core application scenario of sCT, is the most direct index to evaluate its clinical applicability. Figures 10 and 11 show the comparative analysis results of sCT and real CT generated by this method under different dose conditions. Taking the case of brain tumor as an example Fig. 10, the difference in dose distribution between the two scenarios is presented on the left side, and the corresponding DVH comparisons are given on the right side (the solid line represents the real CT, and the dotted line represents the method of this paper). The results show that the dose distributions of the generated sCT are highly consistent with the clinical protocols, especially in terms of the prescribed dose in the CTV region and the conformality of the 75% and 50% isodose lines, which show excellent matching results. For the H&N tumor cases Fig. 11, the analysis also focused on the dose distribution characteristics of the neck metastases. The three-dimensional dose distribution maps and DVH curves showed that the differences at all dose levels were within acceptable limits, which strongly confirmed the clinical applicability and robustness of the method at different anatomical sites.

Model interpretability

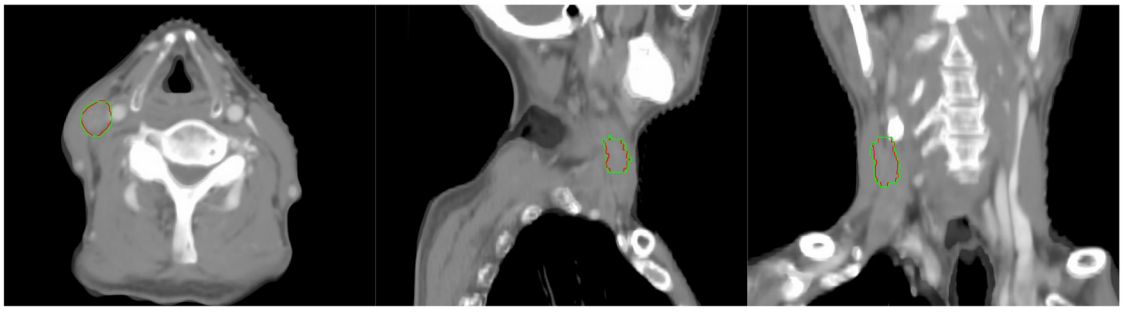
The interpretability of the model aids in understanding the process of generating sCT images and supports the improvement of model performance, enhancing its trustworthiness, transparency, and compliance, thereby increasing its reliability and acceptability in clinical applications. In high-risk fields such as medicine, good

Readers	Evaluation metrics	Image quality score			*P* value		
		CBCT	CT	Ours	CBCTvs. CT	CBCTvs. ours	CT vs. ours
Reader1	Bone structure clarity	2.04 ± 0.715	4.16 ± 0.711	4.15 ± 0.802	<0.001	<0.001	0.771
Reader2		2.01 ± 0.735	3.96 ± 0.752	3.85 ± 0.792	<0.001	<0.001	0.183
Reader1	Detail of local organ segmentation	1.97 ± 0.788	3.84 ± 0.863	3.83 ± 0.742	<0.001	<0.001	0.443
Reader2		1.94 ± 0.742	3.93 ± 0.751	3.93 ± 0.859	<0.001	<0.001	0.701
Reader1	Image deformation distortion	2.02 ± 0.817	3.91 ± 0.788	3.90 ± 0.820	<0.001	<0.001	0.517
Reader2		2.08 ± 0.744	4.00 ± 0.773	3.98 ± 0.712	<0.001	<0.001	0.396
Reader1	Overall image quality	2.27 ± 0.751	4.22 ± 0.738	4.20 ± 0.733	<0.001	<0.001	0.384
Reader2		2.09 ± 0.843	4.13 ± 0.767	4.11 ± 0.833	<0.001	<0.001	0.443

Table 4. Subjective quality ratings of the three sets of images by two reviewers.



(a) Brain tumor target area outlining. From left to right, representing transverse, sagittal and coronal planes, red solid line is real CT, yellow solid line is synthetic CT.



(b) H&N tumor target area outlining. From left to right, representing transverse, sagittal and coronal planes, red solid line is real CT, green solid line is synthetic CT.

Fig. 7. Visualization of tumor target area outlining in Brain and H&N.

interpretability helps reduce misdiagnoses and treatment deviations, promoting the healthy development of artificial intelligence in the medical domain. This study selected the IB-MLP module at the bottom of the U-Net, the ViT module in the decoder, and the convolutional module adjacent to the ViT in the decoder. The noise reduction effects of these three modules were visualized and presented. In Fig. 12, the original image is an unprocessed CBCT scan, heavily contaminated with noise and artifacts, which leads to blurred details and hinders effective clinical diagnosis. After IB-MLP processing, noise is significantly reduced in the image, especially in the brain region (indicated by green arrows). However, some detail loss is still visible in the central part of the image (indicated by the yellow circle), and some residual noise is present in the edge regions, indicating IB-MLP's limitations in detail restoration. IB-MLP focuses primarily on compressing input data and removing irrelevant information, thereby improving information processing efficiency, but it lacks effectiveness in recovering high-frequency details. After ViT processing, the global features of the image are significantly improved. ViT leverages its self-attention mechanism to focus on features from different regions of the image, enhancing the overall understanding of the image, particularly in detail restoration and noise suppression. The regions marked by orange arrows demonstrate that ViT significantly improves structural details, such as the brain contour. Compared to IB-MLP, ViT more effectively restores global features and details, particularly improving the boundary clarity in the brain region, resulting in significantly better image quality than the IB-MLP processed image. In contrast to ViT and IB-MLP, CNN relies more on local feature extraction during image processing. The image processed by CNN shows clearer details in certain areas (such as those indicated by yellow arrows and red ellipses), suggesting that CNN is better suited for processing local features. Through the visualization analysis of the processing results from different modules, we have validated the effectiveness of the HUDiff model in the task of CBCT-to-CT image synthesis. HUDiff combines the advantages of information bottleneck theory, ViT's self-attention mechanism, and convolution, effectively removing noise while restoring both global and local details, providing more accurate and clearer images for medical image synthesis. This is particularly valuable for complex tumor treatment planning and organ localization tasks, with significant clinical implications.

Model complexity

In image synthesis tasks, the computational complexity of models plays a crucial role in determining their application efficiency and feasibility. Table 5 presents a comparison of the computational complexity of various models in terms of average training time per image, inference time per image, parameter count, and memory consumption. The CycleGAN series (including CycleGAN, ADCycleGAN, and IViTCycleGAN), which adopts a GAN architecture, is characterized by relatively fast inference, with inference times shorter than those of models based on diffusion processes. However, these models tend to have higher parameter counts and memory

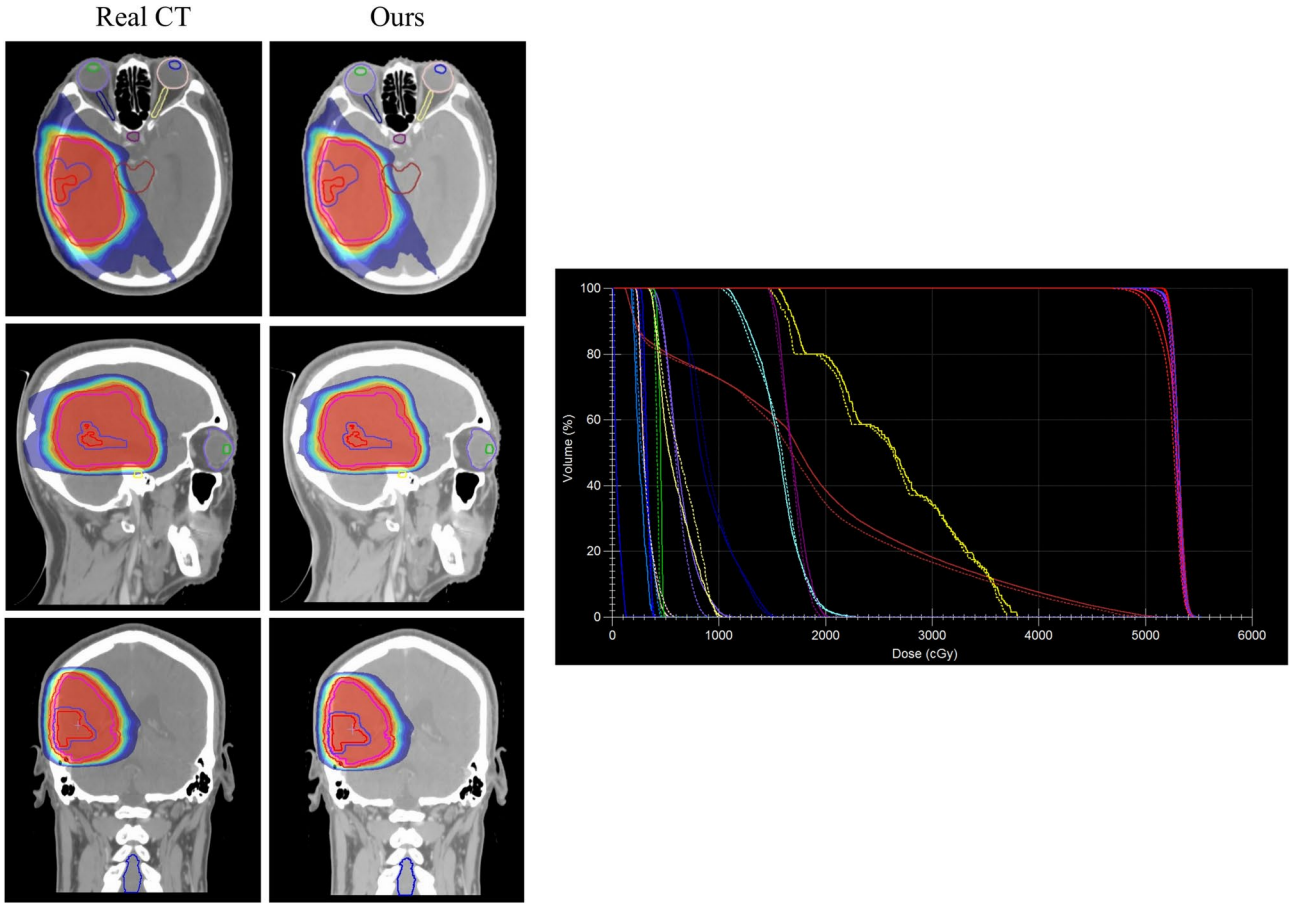


Fig. 8. Spatial distribution analysis of Hounsfield units across the Brain dataset validation series.

consumption, especially IViTCycleGAN, which integrates the ViT module, leading to a significant increase in both parameters and memory requirements. In contrast, diffusion-based models such as DDPM and its variants, EGDiff and HUDiff, demonstrate greater efficiency in terms of parameter count and memory consumption. Specifically, EGDiff optimizes the training process by randomly selecting time steps in each iteration, effectively reducing training time and enhancing training efficiency. Although the inference time for DDPM and EGDiff is longer compared to the CycleGAN series, their parameter count and memory consumption are significantly lower. EGDiff not only maintains the inference time advantage of DDPM but also improves the quality of generated images, further boosting its generation efficiency under comparable computational complexity. HUDiff, building upon DDPM, introduces the ViT module, the IB-MLP module, and dynamic modulation factors, which notably enhance the quality of generated images in terms of both fine details and global consistency. While the inclusion of these modules results in a modest increase in training time, inference time, parameter count, and memory consumption, the improvements in generation quality are substantial. Through a multi-faceted experimental analysis, including visual demonstrations, subjective assessments by experienced clinicians, and a comprehensive evaluation using various quantitative metrics, HUDiff has demonstrated the best performance in terms of generation quality. In conclusion, HUDiff exhibits a significant advantage in balancing computational complexity with generation quality, making it particularly well-suited for applications that require high-quality image synthesis. This model offers an effective solution that achieves an optimal trade-off between generation efficiency and quality in image synthesis tasks.

Discussion

This study presents a novel hybrid U-Net conditional diffusion model framework, HUDiff, designed for generating high-quality sCT images from CBCT data. We conducted a comprehensive analysis using quantitative metrics such as MAE, PSNR, SSIM, and NCC, alongside visual assessments and HU values. The results indicate that the sCT images generated by HUDiff achieve optimal performance compared to the latest algorithms.

Techniques for synthesizing CT images from CBCT data have emerged as a significant research area. This approach not only leverages the benefits of conventional CT imaging but also effectively mitigates the inherent technical limitations of CBCT. By employing advanced image synthesis techniques, the sCT has notably enhanced image quality, tissue contrast, and the integrity of anatomical structures, thereby laying a solid groundwork for precise diagnosis in clinical radiotherapy and the customization of treatment strategies. Initially, GANs were the prevailing method in this domain. GANs achieve image synthesis through a dynamic interaction between

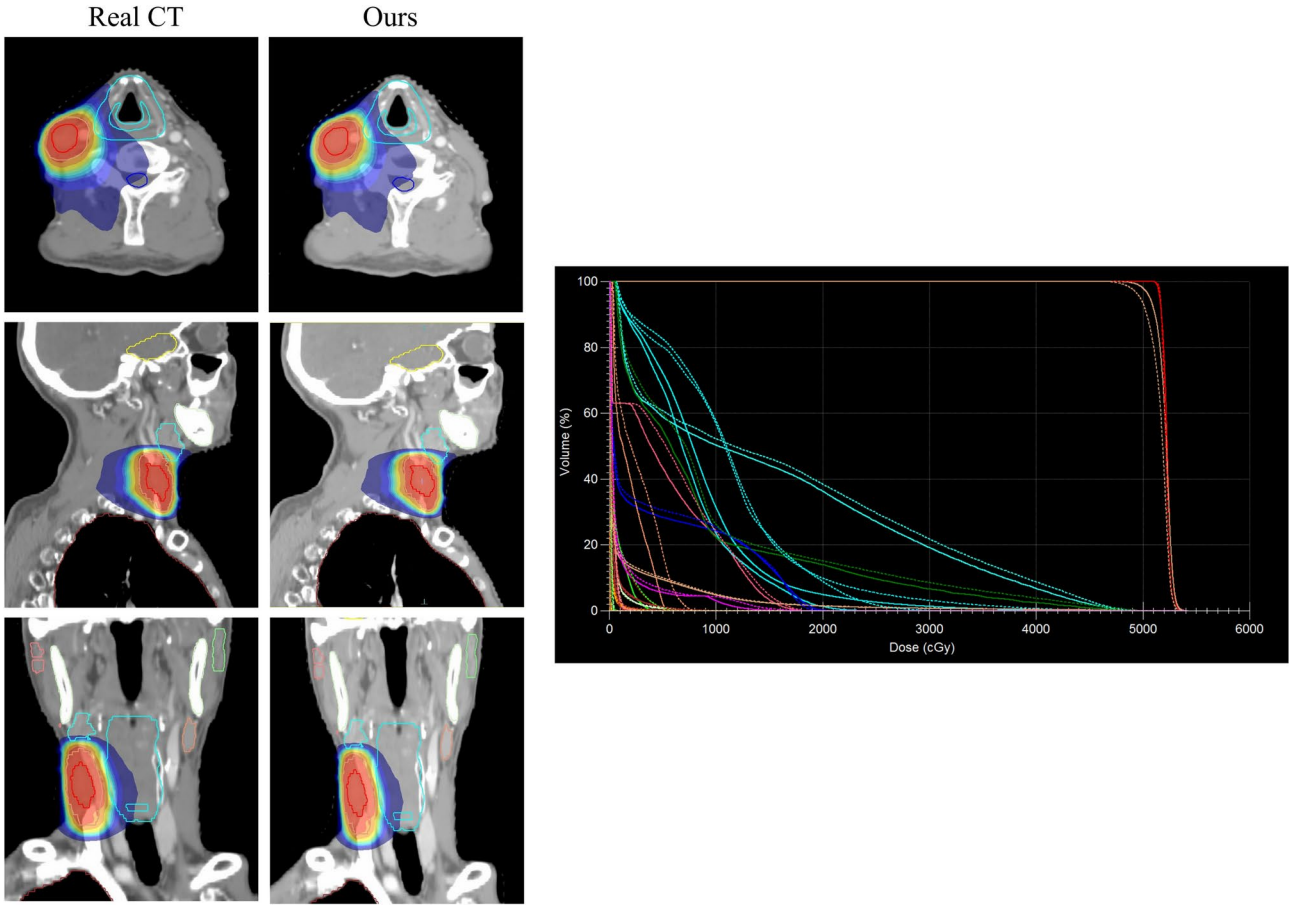


Fig. 9. Spatial distribution analysis of Hounsfield units across the H&N dataset validation series.

the generator and the discriminator, with the former aiming to produce realistic synthetic images. However, this adversarial training approach has inherent drawbacks: an unbalanced capability between the generator and discriminator can lead to mode collapse, which not only diminishes the variety of the generated images but also diminishes the model's interpretability. To address the above challenges, researchers have introduced unsupervised or self-supervised diffusion-based models into the field of sCT synthesis. Compared to GANs, diffusion models provide a more stable training environment and a clearer generation process, indicating their potential for generating high-quality sCT images. However, traditional diffusion models still face three major challenges during the reverse denoising process: (1) difficulty in effectively capturing the global semantic information of the image, (2) potential loss of critical local details during denoising, and (3) the challenge of guiding the model to focus on key features of lesion regions in the absence of paired data, ensuring anatomical consistency in the synthesized images.

In contrast to traditional diffusion models, we propose a hybrid U-Net diffusion model based on ViT and information bottleneck theory. Specifically, we first addressed the limitations of the original U-Net, which could only retain and propagate local feature information, by developing a ViT-based U-Net framework that utilizes self-attention mechanisms. This allows the model to automatically focus on information from various locations in the image, enhancing the understanding of global features. Secondly, we introduced a VIB module at the bottom of the U-Net to filter out redundant and irrelevant information and compress key input data, effectively summarizing the input and better capturing essential features. Lastly, we implemented dynamic modulation factors to balance the contributions of features from the backbone network and skip connections, thereby enhancing reverse denoising in conjunction with the diffusion model. To thoroughly evaluate our model, we performed multi-faceted validation on two independent datasets. As shown in Table 1, the DDPM and its improved algorithm EGDiff outperformed CycleGAN and its variations in all evaluation metrics, demonstrating the advantages of DDPM in generating sCT images. Furthermore, our method surpassed both DDPM and its improved variants, indicating the effectiveness of our designed denoising network architecture. Table 2 illustrates that our approach exhibits significant differences compared to other algorithms across both datasets, further emphasizing its advancement and versatility. Additionally, we conducted extensive ablation studies, revealing that each module contributes positively to overall performance. Notably, ViT outperformed the IB-MLP and DMF in enhancing model performance. Beyond quantitative analysis, we provided visual demonstrations of the sCT images generated by our method. Figures 3 and 4 show that CycleGAN and its improved algorithms produced generation errors and artifacts in the regions of interest. Although DDPM and EGDiff have some

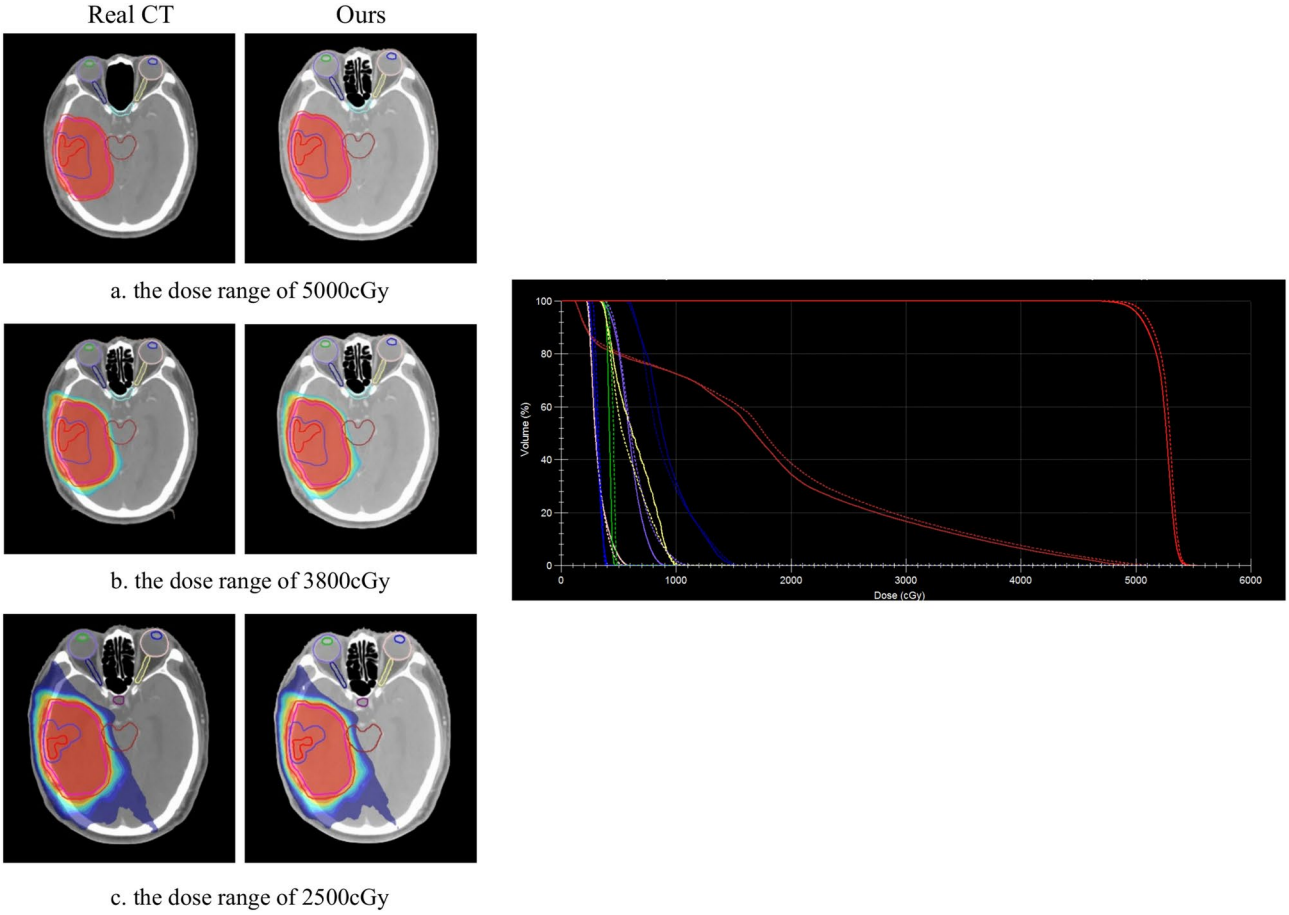


Fig. 10. Voxel-wise analysis of CT attenuation distributions in the Brain dataset validation cohort.

advantages in detail generation compared to CycleGAN, they still exhibit significant detail loss and boundary blurriness. In contrast, our method demonstrated clear superiority over both DDPM and EGDiff, effectively retaining key information by automatically focusing on various locations within the image Fig. 5a and b further validate the advantages of the sCT images generated by our method from the perspective of error maps. Moreover, as shown in Fig. 6a and b, the HU values of our method align closely with the trends of real CT values, indicating that our generated sCT images are more accurate, with minimal deviation from actual CT scans. In addition to the objective index evaluation, we also performed a subjective evaluation, and the subjective quality scores of the three groups of images by the two reviewers are shown in Table 4. CT and sCT had higher scores than CBCT in terms of bony structure clarity, local organ segmentation details, image distortion distortion, and overall image quality scores, and the difference was statistically significant (P <0.001). While the difference between CT and sCT in terms of scores was not statistically significant (P>0.05). This indicates that sCT is close to the reference image CT in terms of subjective clinical scoring and has a very high degree of authenticity. These multi-dimensional experimental results consistently show that the hybrid architecture proposed in this paper can effectively enhance the generation quality of sCT images, and achieve significant improvements in detail preservation, noise suppression and physical property fidelity.

In future research, we plan to further expand HUDiff and explore its potential applications in tumor target delineation. Traditional delineation methods are often challenged by the complexities of tumor structures and surrounding tissues. The high-quality sCT images generated by HUDiff are expected to provide more accurate anatomical information, thus optimizing delineation precision and offering a more reliable basis for target definition in adaptive radiotherapy. Furthermore, investigating how to optimize model architecture and algorithms to reduce computational demands while enhancing the efficiency of diffusion models will also be an important area of research. Additionally, due to the superior spatial information preservation offered by 3D images compared to 2D, exploring diffusion models based on 3D imaging will contribute to the development of more efficient and precise medical image analysis tools, providing reliable technical support for personalized treatment planning.

Conclusions

This study presents HUDiff, an innovative hybrid U-Net architecture integrated with conditional diffusion mechanisms for translating CBCT data into diagnostic-quality synthetic CT representations. Comprehensive evaluations on the Brain and H&N datasets, encompassing both quantitative metrics and perceptual assessments,

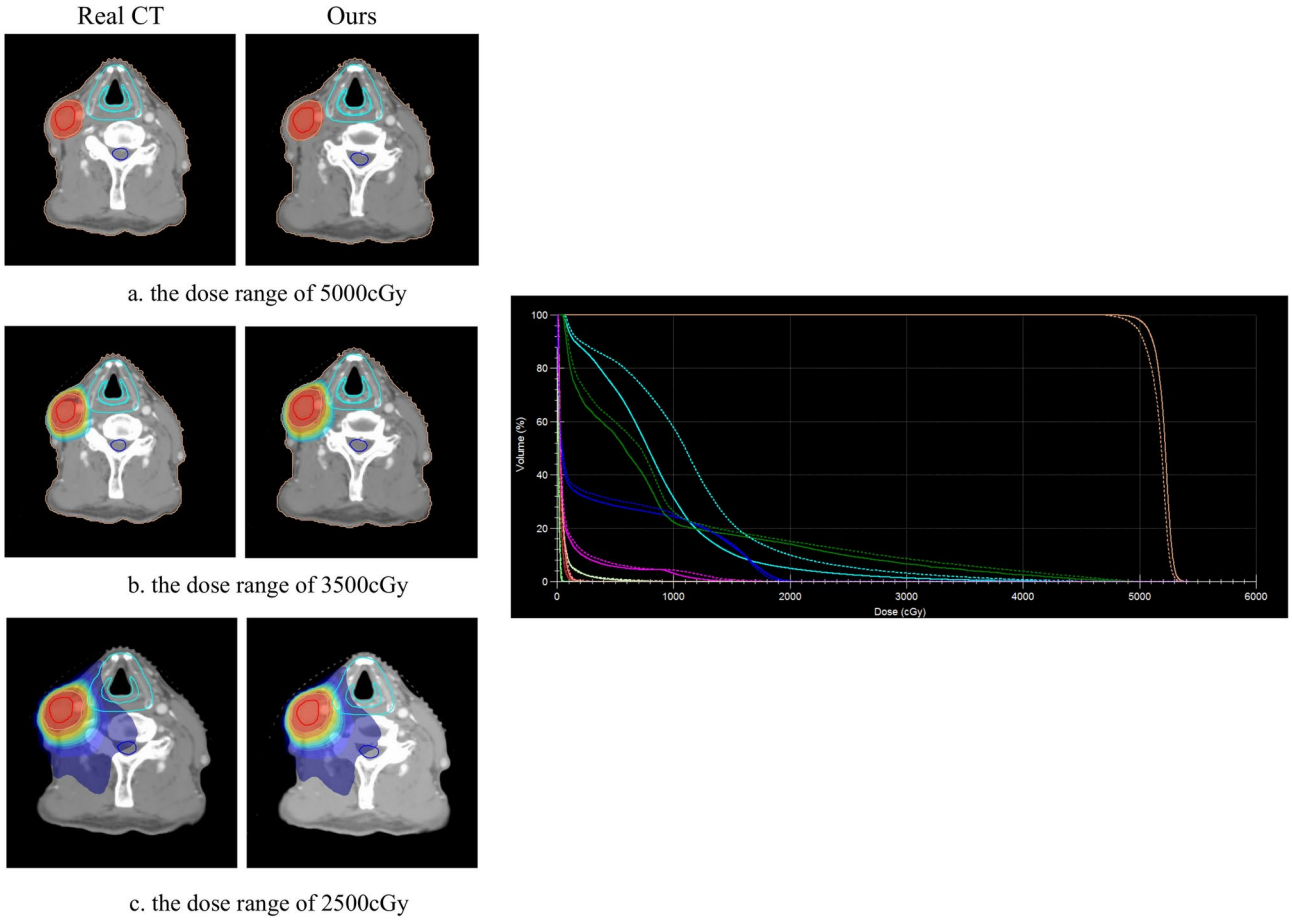


Fig. 11. Voxel-wise analysis of CT attenuation distributions in the H&N dataset validation cohort.

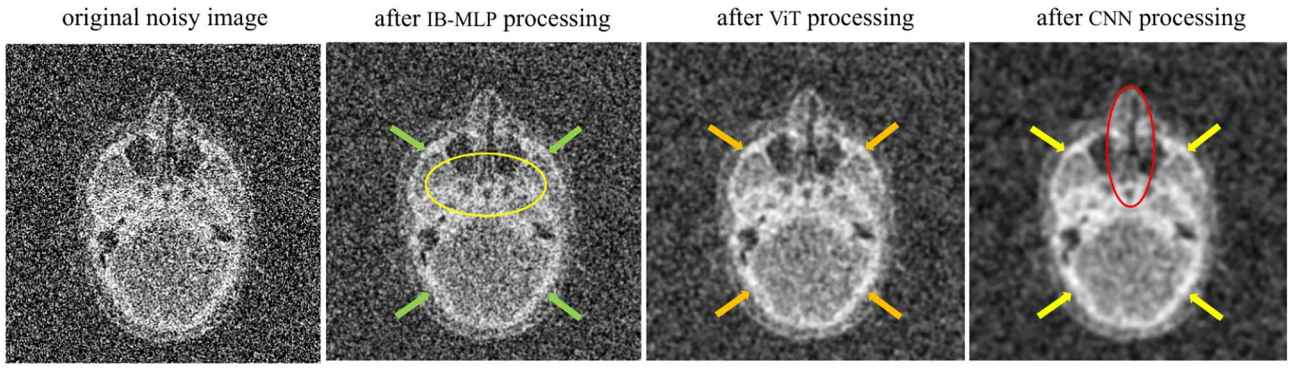


Fig. 12. Visualization of different modules for processing noisy images.

	CycleGAN	ADCycleGAN	IViTCycleGAN	DDPM	EGDiff	HUDiff
Training (s)	0.0061	0.0055	0.0067	0.0036	0.0036	0.0038
Inference (s)	0.0074	0.0084	0.0093	0.2166	0.2166	0.2171
Parameters (M)	31.56	32.54	37.97	24.31	24.31	25.14
Memory (GB)	3.87	3.99	4.29	2.86	2.86	2.93

Table 5. Quantification of computational complexity for different models.

demonstrate HUDiff's superior performance compared to existing approaches. The framework exhibits robust generalization capabilities and enhanced accuracy in CT synthesis, validating its potential for clinical implementation. Performance analysis confirms the system's reliability in supporting radiation oncologists with treatment planning optimization. Finally, integrating frequency-domain information or energy-guided mechanisms may enable unsupervised diffusion models to generate high-quality sCT images without the need for paired data. Additionally, exploring semi-supervised learning to further reduce data requirements while preserving the anatomical consistency of synthesized images remains an important direction for our future research.

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Received: 16 November 2024; Accepted: 25 February 2025
Published online: 28 March 2025

References

1. Jaffray, D. A. Image-guided radiotherapy: From current concept to future perspectives. *Nat. Rev. Clin. Oncol.* **9**, 688–699. https://doi.org/10.1038/nrclinonc.2012.194 (2012).
2. Dawson, L. A. & Sharpe, M. B. Image-guided radiotherapy: Rationale, benefits, and limitations. *Lancet Oncol.* **7**, 848–858. https://doi.org/10.1016/s1470-2045(06)70904-4 (2006).
3. Schulze, R. et al. Artefacts in CBCT: A review. *Dentomaxillofac. Radiol.* **40**, 265–273. https://doi.org/10.1259/dmfr/30642039 (2011).
4. Cho, P. S., Johnson, R. H. & Griffin, T. W. Cone-beam CT for radiotherapy applications. *Phys. Med. Biol.* **40**, 1863. https://doi.org/10.1088/0031-9155/40/11/007 (1995).
5. Barrett, J. F. & Keat, N. Artifacts in CT: Recognition and avoidance. *Radiographics* **24**, 1679–1691. https://doi.org/10.1148/rg.246045065 (2004).
6. Siewerdsen, J. H., Moseley, D., Bakhtiar, B., Richard, S. & Jaffray, D. A. The influence of antiscatter grids on soft-tissue detectability in cone-beam computed tomography with flat-panel detectors: Antiscatter grids in cone-beam CT. *Med. Phys.* **31**, 3506–3520. https://doi.org/10.1118/1.1819789 (2004).
7. Cai, W., Ning, R. & Conover, D. Scatter correction using beam stop array algorithm for cone-beam CT breast imaging. In *Medical Imaging 2006: Physics of Medical Imaging*, Vol. 6142, 1157–1165 (SPIE, 2006). https://doi.org/10.1117/12.655587.
8. Bechara, B., Moore, W., McMahan, C. & Noujeim, M. Metal artefact reduction with cone beam CT: An in vitro study. *Dentomaxillofac. Radiol.* **41**, 248–253. https://doi.org/10.1259/dmfr/80899839 (2012).
9. Li, Y., Garrett, J. & Chen, G.-H. Reduction of beam hardening artifacts in cone-beam ct imaging via smart-recon algorithm. In *Proceedings of SPIE–the International Society for Optical Engineering*, Vol. 9783. https://doi.org/10.1117/12.2216882 (NIH Public Access, 2016).
10. Xie, S., Zhuang, W. & Li, H. An energy minimization method for the correction of cupping artifacts in cone-beam CT. *J. Appl. Clin. Med. Phys.* **17**, 307–319. https://doi.org/10.1120/jacmp.v17i4.6023 (2016).
11. Sun, M. & Star-Lack, J. Improved scatter correction using adaptive scatter kernel superposition. *Phys. Med. Biol.* **55**, 6695. https://doi.org/10.1088/0031-9155/55/22/007 (2010).
12. Gao, L. et al. Streaking artifact reduction for CBCT-based synthetic CT generation in adaptive radiotherapy. *Med. Phys.* **50**, 879–893. https://doi.org/10.1002/mp.16017 (2023).
13. Nomura, Y., Xu, Q., Shirato, H., Shimizu, S. & Xing, L. Projection-domain scatter correction for cone beam computed tomography using a residual convolutional neural network. *Med. Phys.* **46**, 3142–3155. https://doi.org/10.1002/mp.13583 (2019).
14. Jiang, Y. et al. A generalized image quality improvement strategy of cone-beam CT using multiple spectral CT labels in pix2pix GAN. *Phys. Med. Biol.* **67**, 115003. https://doi.org/10.1088/1361-6560/ac6bda (2022).
15. Veldkamp, W. J., Thijssen, M. A. & Karssemeijer, N. The value of scatter removal by a grid in full field digital mammography. *Med. Phys.* **30**, 1712–1718. https://doi.org/10.1118/1.1584044 (2003).
16. Sidky, E. Y., Kao, C.-M. & Pan, X. Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT. *J. X-ray Sci. Technol.* **14**, 119–139 (2006).
17. Xu, Y. et al. A practical cone-beam CT scatter correction method with optimized Monte Carlo simulations for image-guided radiation therapy. *Phys. Med. Biol.* **60**, 3567. https://doi.org/10.1088/0031-9155/60/9/3567 (2015).
18. Usui, K. et al. SU-F-J-70: Monte Carlo study on a cone-beam computed tomography using a cross-type carbon fiber antiscatter grid. *Med. Phys.* **43**, 3422–3422. https://doi.org/10.1118/1.4955978 (2016).
19. Zhu, L., Xie, Y., Wang, J. & Xing, L. Scatter correction for cone-beam CT in radiation therapy. *Med. Phys.* **36**, 2258–2268. https://doi.org/10.1118/1.3130047 (2009).
20. Stankovic, U., Ploeger, L. S., van Herk, M. & Sonke, J.-J. Optimal combination of anti-scatter grids and software correction for CBCT imaging. *Med. Phys.* **44**, 4437–4451. https://doi.org/10.1002/mp.12385 (2017).
21. Wang, L. et al. Automated segmentation of dental CBCT image with prior-guided sequential random forests. *Med. Phys.* **43**, 336–346. https://doi.org/10.1118/1.4938267 (2016).
22. Kida, S. et al. Visual enhancement of cone-beam CT by use of CycleGAN. *Med. Phys.* **47**, 998–1010. https://doi.org/10.1002/mp.13963 (2020).
23. Chen, L., Liang, X., Shen, C., Jiang, S. & Wang, J. Synthetic CT generation from CBCT images via deep learning. *Med. Phys.* **47**, 1115–1125. https://doi.org/10.1002/mp.13978 (2020).
24. Chen, L. et al. Synthetic CT generation from CBCT images via unsupervised deep learning. *Phys. Med. Biol.* **66**, 115019. https://doi.org/10.1088/1361-6560/ac01b6 (2021).
25. Harms, J. et al. Paired cycle-GAN-based image correction for quantitative cone-beam computed tomography. *Med. Phys.* **46**, 3998–4009. https://doi.org/10.1002/mp.13656 (2019).
26. Zhang, Y. et al. Generating synthesized computed tomography from CBCT using a conditional generative adversarial network for head and neck cancer patients. *Technol. Cancer Res. Treat.* **21**, 15330338221085358. https://doi.org/10.1177/15330338221085358 (2022).
27. Liang, J. et al. Sketch guided and progressive growing GAN for realistic and editable ultrasound image synthesis. *Med. Image Anal.* **79**, 102461. https://doi.org/10.1016/j.media.2022.102461 (2022).
28. Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232. https://doi.org/10.1109/iccv.2017.244 (2017).
29. Liu, Y. et al. CBCT-based synthetic CT generation using deep-attention cycleGAN for pancreatic adaptive radiotherapy. *Med. Phys.* **47**, 2472–2483. https://doi.org/10.1002/mp.14121 (2020).

30. Liang, X. et al. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using cycleGAN for adaptive radiation therapy. *Phys. Med. Biol.* **64**, 125002. https://doi.org/10.1088/1361-6560/ab22f9 (2019).
31. Sun, H. et al. Imaging study of pseudo-CT synthesized from cone-beam CT based on 3d cycleGAN in radiotherapy. *Front. Oncol.* **11**, 603844. https://doi.org/10.3389/fonc.2021.603844 (2021).
32. Deng, L., Hu, J., Wang, J., Huang, S. & Yang, X. Synthetic CT generation based on CBCT using respath-cycleGAN. *Med. Phys.* **49**, 5317–5329. https://doi.org/10.1002/mp.15684 (2022).
33. Zhang, X. et al. Combining physics-based models with deep learning image synthesis and uncertainty in intraoperative cone-beam CT of the brain. *Med. Phys.* **50**, 2607–2624. https://doi.org/10.1002/mp.16351 (2023).
34. Li, Z. et al. Using RegGAN to generate synthetic CT images from CBCT images acquired with different linear accelerators. *BMC Cancer* **23**, 828. https://doi.org/10.1186/s12885-023-11274-7 (2023).
35. Szmul, A. et al. Deep learning based synthetic CT from cone beam CT generation for abdominal paediatric radiotherapy. *Phys. Med. Biol.* **68**, 105006. https://doi.org/10.1088/1361-6560/acc921 (2023).
36. Liu, Y. et al. Ct synthesis from CBCT using a sequence-aware contrastive generative network. *Comput. Med. Imaging Graph.* **109**, 102300. https://doi.org/10.1016/j.compmedimag.2023.102300 (2023).
37. Hu, Y., Zhou, H., Cao, N., Li, C. & Hu, C. Synthetic CT generation based on CBCT using improved vision transformer cycleGAN. *Sci. Rep.* **14**, 11455. https://doi.org/10.1038/s41598-024-61492-7 (2024).
38. Sun, H. et al. Research on new treatment mode of radiotherapy based on pseudo-medical images. *Comput. Methods Programs Biomed.* **221**, 106932. https://doi.org/10.1016/j.cmpb.2022.106932 (2022).
39. Zhang, Y. et al. Breath-hold CBCT-guided CBCT-to-CT synthesis via multimodal unsupervised representation disentanglement learning. *IEEE Trans. Med. Imaging* **42**, 2313–2324. https://doi.org/10.1109/tmi.2023.3247759 (2023).
40. Fu, L. et al. Energy-guided diffusion model for CBCT-to-CT synthesis. *Comput. Med. Imaging Graph.* **113**, 102344. https://doi.org/10.1016/j.compmedimag.2024.102344 (2024).
41. Peng, J. et al. CBCT-based synthetic CT image generation using conditional denoising diffusion probabilistic model. *Med. Phys.* **51**, 1847–1859. https://doi.org/10.1002/mp.16704 (2024).
42. Chen, X. et al. CBCT-based synthetic CT image generation using a diffusion model for CBCT-guided lung radiotherapy. *Med. Phys.*[SPACE]https://doi.org/10.1002/mp.17328 (2024).
43. Zhang, Y. et al. Texture-preserving diffusion model for CBCT-to-CT synthesis. *Med. Image Anal.*[SPACE]https://doi.org/10.1016/j.media.2024.103362 (2024).
44. Saharia, C. *et al.* Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10. https://doi.org/10.1145/3528233.3530757 (2022).
45. Cao, H. et al. A survey on generative diffusion models. *IEEE Trans. Knowl. Data Eng.*[SPACE]https://doi.org/10.1109/tkde.2024.3361474 (2024).
46. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695. https://doi.org/10.1109/cvpr52688.2022.01042 (2022).
47. Saharia, C. et al. Image super-resolution via iterative refinement. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**, 4713–4726. https://doi.org/10.1109/tpami.2022.3204461 (2022).
48. Li, Y. et al. Zero-shot medical image translation via frequency-guided diffusion models. *IEEE Trans. Med. Imaging*[SPACE]https://doi.org/10.1109/TMI.2023.3325703 (2023).
49. Azad, R. et al. Advances in medical image analysis with vision transformers: A comprehensive review. *Med. Image Anal.* **91**, 103000. https://doi.org/10.1016/j.media.2023.103000 (2024).
50. Shamshad, F. et al. Transformers in medical imaging: A survey. *Med. Image Anal.* **88**, 102802. https://doi.org/10.1016/j.media.2023.102802 (2023).
51. Manzari, O. N., Ahmadabadi, H., Kashiani, H., Shokouhi, S. B. & Ayatollahi, A. Medvit: A robust vision transformer for generalized medical image classification. *Comput. Biol. Med.* **157**, 106791. https://doi.org/10.1016/j.compbiomed.2023.106791 (2023).
52. Dalmaz, O., Yurt, M. & Çukur, T. Resvit: Residual vision transformers for multimodal medical image synthesis. *IEEE Trans. Med. Imaging* **41**, 2598–2614. https://doi.org/10.1109/TMI.2022.3167808 (2022).
53. He, A. et al. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *IEEE Trans. Med. Imaging* **42**, 2763–2775. https://doi.org/10.1109/TMI.2023.3264513 (2023).
54. Hu, S., Lou, Z., Yan, X. & Ye, Y. A survey on information bottleneck. *IEEE Trans. Pattern Anal. Mach. Intell.*[SPACE]https://doi.org/10.1109/TPAMI.2024.3366349 (2024).
55. Li, G., Jin, D., Yu, Q. & Qi, M. IB-TransUNet: Combining information bottleneck and transformer for medical image segmentation. *J. King Saud Univ. Comput. Inf. Sci.* **35**, 249–258. https://doi.org/10.1016/j.jksuci.2023.02.012 (2023).
56. Li, G., Zheng, Y., Cui, J., Gai, W. & Qi, M. DIM-UNet: Boosting medical image segmentation via diffusion models and information bottleneck theory mixed with mlp. *Biomed. Signal Process. Control* **91**, 106026. https://doi.org/10.1016/j.bspc.2024.106026 (2024).
57. Si, C., Huang, Z., Jiang, Y. & Liu, Z. Freeu: Free lunch in diffusion U-Net. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4733–4743. https://doi.org/10.1109/cvpr52733.2024.00453 (2024).
58. Wang, J., Wu, Q. J. & Pourpanah, F. An attentive-based generative model for medical image synthesis. *Int. J. Mach. Learn. Cybern.* **14**, 3897–3910. https://doi.org/10.1007/s13042-023-01871-0 (2023).

Acknowledgements

This research was funded by the Jiangsu Provincial Key Research and Development Program (BE2020714).

Author contributions

Conceptualization, H.Z., X.L. and C.H.; methodology, C.H.; software, C.H.; validation, C.H. and N.C.; formal analysis C.H. and N.C.; resources C.H. and N.C; writing—original draft preparation, C.H.; writing—review and editing, C.H., Y.H.; funding acquisition, N.C. All authors have read and agreed to the published version of the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Ethical statement

The study was approved by the Research Ethics Committee of Nanjing Medical University (NMUE2021301) and individual informed consent for this retrospective analysis was waived.

Additional information

Correspondence and requests for materials should be addressed to H.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.