Radiation Oncology

RESEARCH

Open Access

Diffusion-CSPAM U-Net: A U-Net model integrated hybrid attention mechanism and diffusion model for segmentation of computed tomography images of brain metastases

Yiren Wang[1,2], Zhongjian Wen[1,2], Shuilan Bao[1,2], Delong Huang[3], Youhua Wang[4], Bo Yang[5], Yunfei Li[5], Ping Zhou[2,6*], Huaiwen Zhang[7*] and Haowen Pang[5*]

Abstract

Background Brain metastases are common complications in patients with cancer and significantly affect prognosis and treatment strategies. The accurate segmentation of brain metastases is crucial for effective radiation therapy planning. However, in resource-limited areas, the unavailability of MRI imaging is a significant challenge that necessitates the development of reliable segmentation models for computed tomography images (CT).

Purpose This study aimed to develop and evaluate a Diffusion-CSPAM-U-Net model for the segmentation of brain metastases on CT images and thereby provide a robust tool for radiation oncologists in regions where magnetic resonance imaging (MRI) is not accessible.

Methods The proposed Diffusion-CSPAM-U-Net model integrates diffusion models with channel-spatial-positional attention mechanisms to enhance the segmentation performance. The model was trained and validated on a dataset consisting of CT images from two centers ($n=205$) and ($n=45$). Performance metrics, including the Dice similarity coefficient (DSC), intersection over union (IoU), accuracy, sensitivity, and specificity, were calculated. Additionally, this study compared models proposed for brain metastases of different sizes with those proposed in other studies.

Results The diffusion-CSPAM-U-Net model achieved promising results on the external validation set. Overall average DSC of 79.3% ± 13.3%, IoU of 69.2% ± 13.3%, accuracy of 95.5% ± 11.8%, sensitivity of 80.3% ± 12.1%, specificity of 93.8% ± 14.0%, and HD of 5.606 ± 0.990 mm were measured. These results demonstrate favorable improvements over existing models.

*Correspondence:
Ping Zhou
zhouping11@swmu.edu.cn
Huaiwen Zhang
1761580890@qq.com
Haowen Pang
haowenpang@foxmail.com
Full list of author information is available at the end of the article

Conclusions The diffusion-CSPAM-U-Net model showed promising results in segmenting brain metastases in CT images, particularly in terms of sensitivity and accuracy. The proposed diffusion-CSPAM-U-Net model provides an effective tool for radiation oncologists for the segmentation of brain metastases in CT images.

Keywords Computed tomography, Image procession, Brain metastases, Automated segmentation, Deep learning, Stereotactic radiosurgery

Introduction

Brain metastases are secondary brain tumors that originate from cancer cells that have spread from other parts of the body [1]. Epidemiological studies have indicated that brain metastases occur in 10% to 40% of patients with solid tumors [2]. Clinically, patients with brain metastases may present with various symptoms including headaches, neurological deficits, seizures, and cognitive disturbances, which can severely affect their quality of life and overall prognosis [3].

The clinical management of brain metastases involves a multidisciplinary approach including surgery, radiotherapy, and systemic therapies [4]. Accurate delineation of the gross tumor volume (GTV) is crucial in treatment planning, especially in radiotherapy, where precise targeting of the tumor while sparing healthy brain tissue is essential in maximizing therapeutic efficacy and minimizing adverse effects [5]. Currently, MRI-sim is commonly used to simulate the localization of patients with brain metastases before stereotactic radiosurgery [6]. However, computed tomography (CT) is still used in healthcare settings in developing countries as the standard of care for the simulated localization of brain metastases for radiotherapy [7].

Moreover, the manual segmentation of the GTV from CT images is highly challenging owing to the heterogeneous appearance of brain metastases, the presence of surrounding edema, and variations in image quality [8]. This process is not only labor-intensive and time-consuming but also subject to significant inter- and intra-observer variability, which leads to inconsistent treatment outcomes. Therefore, a method that can assist radiation oncologists in segmentation is required. Given these challenges, current deep-learning methods, such as convolutional neural networks (CNNs), have shown significant promise for medical image segmentation tasks [9]. Advanced segmentation architectures, such as vision transformers, often require large-scale datasets and extensive computational resources to achieve optimal performance [10]. However, in many developing countries, healthcare facilities often lack the capacity to support resource-intensive models, which limits their practical applicability [11]. In contrast, the U-Net architecture has emerged as a highly effective solution for medical image segmentation, particularly in scenarios with limited data availability [12]. The symmetric encoder-decoder structure of U-Net allows it to capture multiscale contextual information efficiently, making it well-suited for training on smaller datasets [13]. Its relatively simple architecture and low computational requirements render it accessible and feasible for use in resource-constrained environments. Despite its advantages, the standard U-Net model still faces challenges in accurately capturing the fine details and boundaries of tumors, especially in complex and noisy CT images [14]. However, this model may not be able to capture complex texture details in CT images. This limitation highlights the need for further enhancements to improve the sensitivity and accuracy of U-Net in segmenting intricate and clinically significant features in medical images.

To address these challenges, we introduce an integrated approach that combines the strengths of diffusion models and an enhanced U-Net architecture. Diffusion models are known for their powerful capabilities in image denoising and generation, and they offer novel solutions by which to improve the quality and robustness of medical image segmentation [15]. The forward noise addition process in diffusion models involves the gradual addition of controlled noise to the input image, which helps capture the underlying structure and important features, even under noisy conditions [16]. This process enhances the ability of the model to retain crucial details and contrast in the images, which provides a richer set of features for the segmentation task. This step helps accumulate essential gradient information, which makes the image structure more pronounced for subsequent processing [17]. The reverse reconstruction process aims to revert a noisy image into its enhanced state [18].

In addition to leveraging diffusion models, this study enhances the U-Net architecture by integrating Channel, Spatial, and Positional Attention Modules (CSPAM). The original Channel Attention Module (CAM) and Spatial Attention Module (SAM) were proposed by Woo et al. [43]. Positional Attention Module (PAM), on the other hand, references the positional attention module in the Dual Attention Network framework proposed by Fu et al. [44]. In this study, we unified all three types of attention modules by integrating them

into the U-Net segmentation framework by means of tandem. It aims to improve the U-Net model's ability to accurately localize segmentation. The purpose of the Channel Attention Module (CAM) is to allocate importance to feature map channels in a CNN. By assigning different weights to each channel, CAM emphasizes the channels that contribute the most to the task while suppressing irrelevant or redundant channels [19]. The Spatial Attention Module (SAM) aims to enhance the feature representation of key regions in an image. Essentially, it transforms spatial information from the original image into another space while retaining critical information. SAM generates weights for each position and applies these weights to the output, thereby enhancing specific regions and simultaneously suppressing irrelevant background areas [45]. PAM calculates the correlation between each position in the feature map and all other positions. Based on this, it generates an attention weight matrix. This matrix is then used to adjust the original feature map, enabling the model to focus more on spatial locations that are highly relevant to the current task. This approach helps capture global contextual information [46].

By combining the noise-resilient capabilities of diffusion models with the enhanced attention mechanisms of U-Net, our integrated approach aims to significantly improve the segmentation accuracy of brain metastases in CT images. This method not only addresses the challenges posed by noisy images but also ensures the efficient utilization of limited data, making it highly suitable for application in resource-constrained healthcare settings. The enhanced U-Net with CSPAM, supported by the robust preprocessing of diffusion models, offers a promising solution for the automatic segmentation of brain metastases. This solution can ultimately aid radiation oncologists in achieving more consistent and precise treatment planning.

Methods

Patients

This study analyzed the data of 250 patients who underwent radiation therapy between January 2016 and January 2022. The patients were divided into two groups: 205 from the Jiangxi Cancer Hospital and 45 from the Affiliated Hospital of Southwest Medical University. The study complied with the Declaration of Helsinki and received ethical approval from the Ethics Review Boards of Jiangxi Cancer Hospital (approval no. 2023KY082) and the Affiliated Hospital of Southwest Medical University (approval no. KY2023041). The need for informed consent was waived due to the retrospective nature of the study. Eligibility criteria included: (1) age 18 or older, (2) availability of comprehensive electronic health and imaging records, and (3) absence of brain abnormalities other than metastases. The exclusion criteria were as follows: (1) imaging records with artifacts or degradation, (2) presence of other brain anomalies, and (3) incomplete patient data. The CT image data used for this study was used in a previous automated segmentation study by our team [20]. The mean age of the patients was 54.352 ± 12.375 and the male/female ratio of patient composition was 50.8% ($n=127$):49.2% ($n=123$). The top three primary sites were lung cancer (60.8%), breast cancer (20%), and melanoma (4.8%) (Supplementary File 1).

Scanning parameters

The CT images from Jiangxi Cancer Hospital were obtained using CT simulation scans for patient positioning during radiotherapy (RT). These scans were performed using a Siemens SOMATOM Definition AS20 Large-bore CT scanner configured with a tube voltage of 120 kVp, tube current of 540 mAs, and scanning range from the skull apex to the third cervical vertebra (C3). The images had a resolution of 512×512 pixels, slice thickness of 3 mm, and field of view (FOV) between 250 and 400 mm. At the Affiliated Hospital of Southwest Medical University, CT scans were acquired for radiotherapy localization using a GE LightSpeed RT 4 scanner with the following settings: 120 kVp tube voltage, 512×512 pixels, slice thickness of 3 mm, and 250–400 mm FOV. Additionally, MRI scans were conducted using a Philips 3.0 T Ingenia MR-sim system that was specifically designed for large-bore magnetic resonance radiation positioning. The sequence for the patient scan was T1W with parameters TR/TE = 5.0/2.4 ms and FOV 280–280 mm.

Preprocessing and segmentation of ground truth

All patient images, including contrast-enhanced CT and MR scans, were resampled to $1 \times 1 \times 1$ mm^3 and processed for denoising, histogram equalization, and grayscale normalization. Normalization was performed with a window width of 100, a window level of 40, and a CT HU range from -10 to 90. The Jiangxi Cancer Hospital dataset was used for model training and internal validation ($n=205$), whereas the Affiliated Hospital of Southwest Medical University dataset was used for external validation ($n=45$). Following contrast enhancement, the CT and MR images were fused. Two radiation oncologists, each with a decade of experience, segmented the gross tumor volume (GTV) of the brain metastases, creating regions of interest (ROIs) that served as ground truth labels. In cases of discrepancies,

a third radiation oncologist with 15 years of experience made the final decision. MR images were used only as references to delineate the ground truth. The automatic segmentation model proposed in this study operates exclusively on CT images.

Diffusion model preprocessing enhancement

Noise addition process

In this study, we adopt a diffusion-model-based preprocessing step inspired by Denoising Diffusion Probabilistic Models (DDPM) [34, 35], which were originally introduced to gradually add noise to an image and later remove that noise to reconstruct the image (Fig. 1). From an initial CT image x_0, we define a sequence of noise level parameters $\{\beta_t\}_{t=1}^{T}$, where each β_t controls the amount of noise added to the image at each time step. For each time step t, the image x_{t-1} is updated to x_t by:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \varepsilon_t\sqrt{\beta_t}, \varepsilon_t \sim N(0,1),$$

where, ε_t is random noise drawn from a standard normal distribution. The values β_t are typically defined within a chosen range $[\beta_{min}, \beta_{max}]$ and can be linearly scheduled across $t = 1, 2, \ldots, T$. For instance, using a linear schedule:

$$\beta_t = \beta_{min} + (\beta_{max} - \beta_{min}) \times \frac{t-1}{T-1}$$

By gradually adding noise, the image transitions from its original state x_0 to a near-pure noise state x_T. This forward process retains the overall structural information of the image while providing learning signal for the subsequent reverse reconstruction phase.

Reverse reconstruction process

After obtaining the high-noise image x_T, the diffusion model employs a reverse (denoising) process to iteratively recover an estimate x_0' of the original image x_0. At each reverse time step t, we generate x_{t-1} conditioned on x_t according to:

$$p(x_{t-1}|x_t) = N(\mu_\theta(x_t,t), \Sigma_\theta(x_t,t)),$$

where $\mu_\theta(x_t,t)$ and $\Sigma_\theta(x_t,t)$ are the conditional mean and conditional variance, respectively. These are learned by a neural network parameterized θ, conditioned on both noisy image x_t d the time step t. We utilize a U-Net architecture, as described in previous studies [36, 37], for predicting $\mu_\theta(x_t,t)$ and $\Sigma_\theta(x_t,t)$. The U-Net is chosen for its robust capability in multi-scale feature extraction and reconstruction, which is particularly advantageous for medical imaging applications [36, 37]. This architecture effectively captures both global context and local details, enabling the network to perform precise denoising. During training, the network learns to minimize the difference between the predicted noise and the actual noise

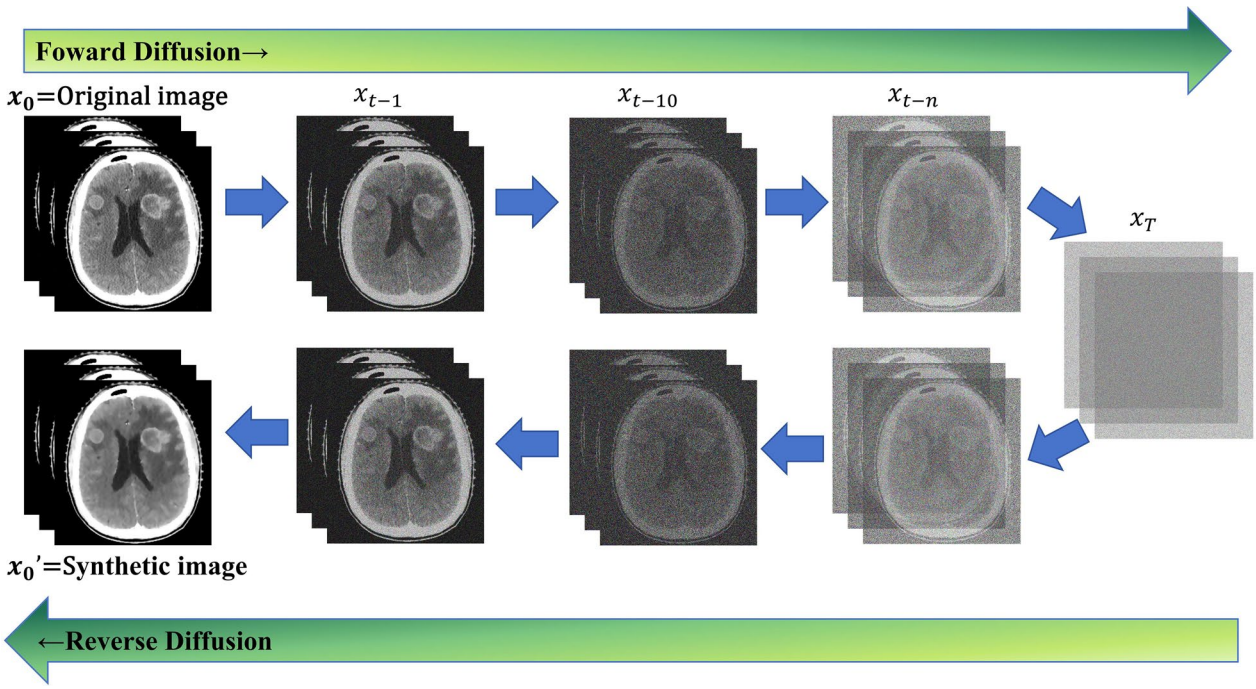

Fig. 1 Diffusion model processing flow schematic

added at each time step. Specifically, the loss function is defined as:

$$L(\theta) = \mathbb{E}_{x_0,t,\varepsilon}\left[\varepsilon - \varepsilon_\theta^2\right],$$

where ε represents the true noise and $\varepsilon_\theta\ (x_t, t)$ is the noise predicted by the network. By optimizing this loss, the U-Net progressively learns to accurately estimate and remove noise from the images across all time steps, thereby enabling effective reconstruction of the original image x_0. Through the above steps, we synthesized the reconstructed CT image through this model, which has similar texture details and structure to enhance the dataset [38, 39].

To evaluate the quality of the reconstructed images x_0' compared to the original images x_0, we employ Peak signal to noise ratio (PSNR) and Structural similarity index (SSIM) to evaluate the synthetic image quality. In addition, we compare the performance with several other classical models for synthesizing CT images.

U-Net architecture with channel, spatial, and positional attention mechanisms

Based on the standard U-Net architecture, we integrated a channel attention module (CAM), spatial attention module (SAM), and positional attention module (PAM) to enhance the network's ability to capture important features and spatial information in images (Fig. 2). In addition, feature space enhancement was achieved by fusing the images processed by the diffusion model with the original images.

The purpose of the channel attention module is to focus on important features in the channel dimensions of the feature map. First, global average pooling and global max pooling are performed on the input feature map K to generate two descriptor vectors: K_{avg} and K_{max}. These two descriptor vectors are passed through the shared, fully connected layers $FC1$ and $FC2$ to generate two attention weight vectors. These attention weight vectors are then added and passed through a sigmoid activation function to generate the channel attention weight J_C:

$$J_C = \sigma\left(FC2\left(\sigma\left(FC1\left(K_{avg}\right)\right)\right) \oplus FC2(\sigma(FC1(K_{max})))\right),$$

where σ represents the sigmoid function.

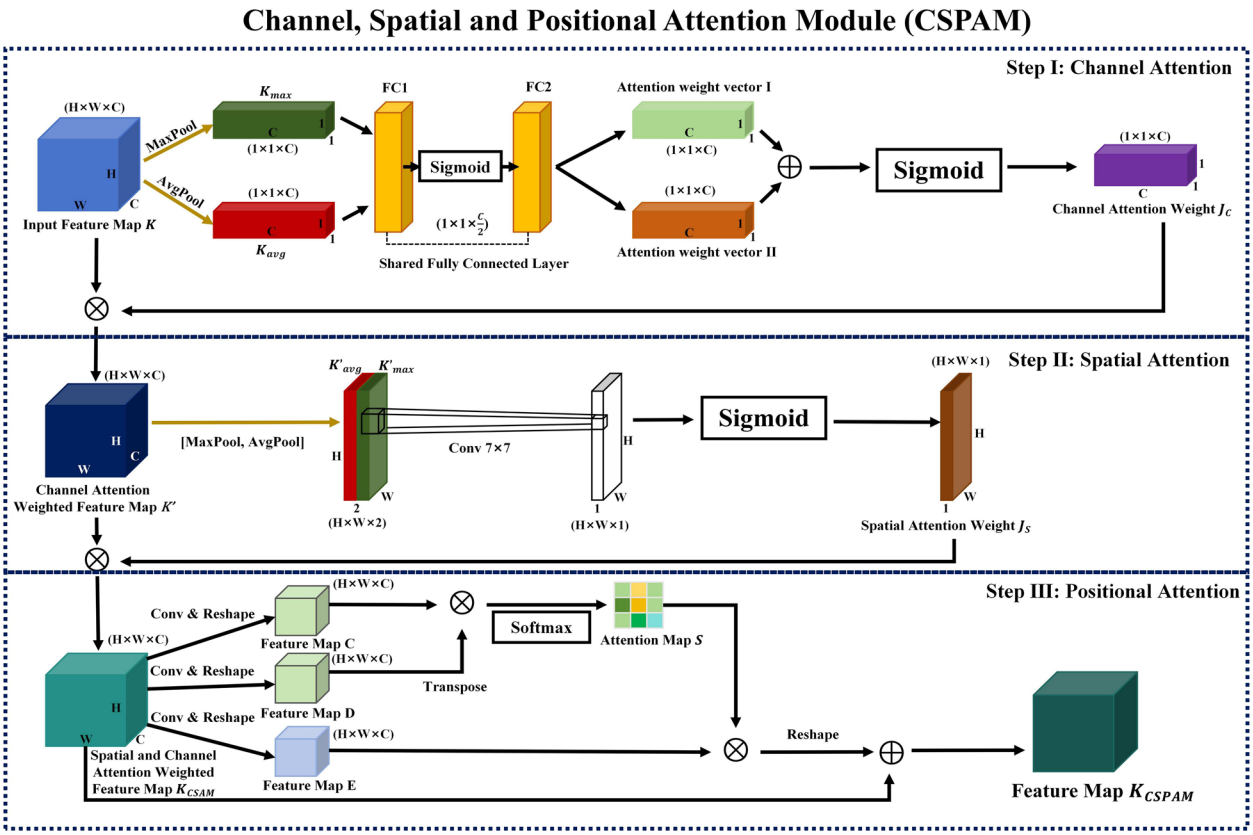

Fig. 2 Channel, spatial, and positional attention module architecture proposed in this study

By element-wise multiplying the channel attention weight J_C with the original feature map K, we obtain the channel-weighted feature map K':

$$K' = J_C \otimes K,$$

The spatial attention module enhances the focus on important regions in the spatial dimension of the feature map. Global average pooling and max pooling operations were performed on the channel-weighted feature map K' along the channel dimension to generate two spatial descriptor vectors: K'_{avg} and K'_{max}. The aggregated feature maps are then passed through a convolution layer, added together, and passed through a sigmoid activation function to generate the spatial attention weight J_S:

$$J_S = \sigma\left(Conv\left(\left[K'_{avg} \oplus K'_{max}\right]\right)\right),$$

By element-wise multiplying the spatial attention weight J_S with the channel-weighted feature map K', we obtain the final spatial-channel attention-weighted feature map K_{CSAM}:

$$K_{CSAM} = J_S \otimes K',$$

The positional attention module aims to enhance the sensitivity of the model to specific spatial information. The input feature map K_{CSAM} was convolved and reshaped into feature maps C, D, and E. A dot product is performed on C and D, and this is followed by a softmax operation to generate the attention map S:

$$S = Softmax\left(C \cdot D^T\right),$$

Then, the dot product of E and S is taken and reshaped into G:

$$G = E \cdot S.$$

Finally, G is added element-wise to the input feature map K_{CSAM} to obtain the spatial channel-positional attention-weighted feature map K_{CSPAM}.

Encoder stage

This study used U-Net as the baseline architecture. Each layer of the encoder consists of two convolutional layers followed by a downsampling operation, with each convolutional layer being followed by a rectified linear unit activation function and a batch normalization layer. The integrated CSPAM module is applied to the output feature map of the last convolutional layer before downsampling. This ensures that the network enhances its sensitivity to the feature channels, spatial regions, and positional information before reducing the feature map size.

Decoder stage

Each layer of the decoder is comprised of two convolutional layers and an upsampling operation. The CSPAM module is inserted after the output of the convolutional layers before upsampling to ensure that the network can effectively capture important feature channels, spatial information, and positional information while restoring the image spatial details. This ultimately enhances the ability of the model to capture specific.

Feature fusion with attention modules

Feature space enhancement is achieved by fusing the features from the original image and the image processed using the diffusion model. The specific processes are as follows:

First, input the original image x_0 and the reconstructed image x'_0 into the improved U-Net model. After processing through the encoder stage with CSPAM, we obtain $K_{CSPAM}(x_0)$ and $K_{CSPAM}(x'_0)$, respectively. The feature maps from the original and enhanced images are then fused using the weight w to adjust their contribution ratios. The fusion formula is:

$$K_{combined} = w(K_{CSPAM}(x_0)) + (1-w)\left(K_{CSPAM}\left(x'_0\right)\right),$$

where $K_{combined}$ represents the final fused feature map, which is then fed into the decoder part of U-Net. In the decoder stage, the features from the original and enhanced images are further integrated using a specially designed feature fusion strategy after each upsampling and a feature fusion operation that is implemented through the CSPAM module. The final feature map is used to generate the segmentation results.

Segmentation boundary refinement

Although the improved U-Net proposed in this study can provide a coarse outline of the tumor region, its segmentation boundaries may still be imprecise or coarse. Therefore, this study introduces a conditional diffusion model to post-process the U-Net segmentation results, aiming to preserve the initial segmentation contours while significantly enhancing the overall shape consistency and boundary precision of the target area. Input image can be defined as:

$$X \in \mathbb{R}^{H \times W \times C}$$

Define the forward inference network based on CSPAM U-Net as:

$$\mathcal{F}_\theta :\in \mathbb{R}^{H \times W \times C} \rightarrow \mathbb{R}^{H \times W}$$

where θ denotes the network parameters. For a given input X, forward inference yields a probability map

$Q = \mathcal{F}_\theta(X)$. Here, $Q(u)$ is the predicted confidence that the pixel u belongs to the target class. To obtain a binary mask for the preliminary segmentation, a threshold $\lambda \in (0,1)$ is introduced, and we define:

$$\tilde{G}(u) = \begin{cases} 1, & if\ Q(u) \geq \lambda \\ 0, & otherwise \end{cases}$$

where the mask $\tilde{G}$, generated by the CSPAM U-Net, may exhibit local deviations in boundary details, partly due to the diverse shapes of brain metastases in CT images, as well as the potential misjudgments of the attention module when dealing with extreme noise or tiny structures. To address this issue, this study incorporates a diffusion model into the post-processing stage and employs a conditional mechanism to make shape refinements more targeted.

The segmentation mask generated by CSPAM U-Net can be defined as $\tilde{G}$. A series of noise-injection steps $\{\gamma_k\}_{k=1}^K$ is established, where each γ_k controls the amount of noise injection into $\tilde{G}$ at the k-th step. Through a forward noise-adding process during training, $\tilde{G}$ is gradually transformed into a distribution approximating isotropic Gaussian noise, which can be expressed as:

$$q(z_k|z_{k-1}) = N\left(z_k;\ \sqrt{\gamma_k}z_{k-1},\ (1-\gamma_k)I\right),\ k = 1, \ldots, K$$

where $z_0 = \tilde{G}$ indicates the segmentation mask generated by CSPAM U-Net, and z_k represents the result after noise injection at the k-th step. I is the identity covariance matrix. In the inference (reverse denoising) phase, the diffusion model needs to progressively restore the random noise $z_k \sim N(0,I)$ back to z_0, which is close to the ground truth shape distribution.

This study was inspired by the study of conditional diffusion model by Rombach et al. [48]. To facilitate the introduction of external information for segmentation refinement, this study adopts a conditional mechanism (*cond*), wherein the ground truth segmentation mask $\tilde{G}_{true}$ and the original image X are jointly provided to guide the denoising process. Formally, this can be expressed as:

$$p_\psi(z_{k-1}|z_k, cond) \approx N\left(z_{k-1};\ \mu_\psi(z_k, cond, k), \Sigma_\psi(z_k, cond, k)\right)$$

where ψ denotes the parameters of the diffusion model, and μ_ψ and Σ_ψ are the mean and covariance output by the conditional neural network. The condition $cond = \left\{\tilde{G}_{true}, X\right\}$ serves as the external prior. Hence at each iteration of k of the reverse process, we have:

$$z_{k-1} = \mu_\psi(z_k, cond, k) + \Sigma_\psi^{\frac{1}{2}}\varepsilon_k, \varepsilon_k \sim N(0,I)$$

Gradually refining the Gaussian noise state z_K back toward the vicinity of the ground truth mask distribution. When the condition includes $\tilde{G}_{true}$, the model retains the primary contour as a "segmentation prior"; original image X provide local gray-level and texture details can be leveraged to enhance the refinement process. This ultimately yields a significantly optimized segmentation result in terms of both overall shape and boundary detail.

Loss function

To optimize the model performance and ensure the accuracy of the segmentation results and boundary refinement, we designed a composite loss function that combines the segmentation and boundary refinement losses. The total loss is composed of the cross-entropy loss L_{seg} and boundary refinement loss L_{refine}, which are expressed as

$$L_{total} = L_{seg}\left(S_{pred}, S_{true}\right) + \lambda L_{refine}\left(S_{refined}, S_{true}\right),$$

where S_{pred} is the predicted segmentation result, S_{true} is the ground truth segmentation label, $S_{refined}$ is the segmentation result after boundary refinement, and λ is the weight coefficient used to balance the two losses. In training the CSPAM U-Net, we employ a combination of Dice loss and cross-entropy loss:

$$L_{seg}\left(S_{pred}, S_{true}\right) = L_{Dice}\left(S_{pred}, S_{true}\right) + L_{CE}\left(S_{pred}, S_{true}\right)$$

The Dice loss is giveby:

$$L_{Dice} = 1 - \frac{2\sum_i \left(S_{pred,i} \times S_{true,i}\right)}{\sum_i S_{pred,i}^2 + \sum_i S_{true,i}^2}$$

where $S_{pred,i}$ and $S_{true,i}$ denote the predicted and ground-truth values at the i-th pixel, respectively. The cross-entropy loss is formulated as:

$$L_{CE} = -\sum_i [S_{true,i} \log\left(S_{pred,i}\right) + \left(1 - S_{true,i}\right)\log\left(1 - S_{pred,i}\right)$$

After obtaining an initial segmentation result from the CSPAM U-Net, a conditional diffusion model is employed to refine the segmentation boundaries. Let ε denote the noise injected into the ground-truth segmentation mask at the k-th diffusion step and let $\varepsilon_\psi(z_k, cond, k)$ be the noise predicted by the diffusion network (with parameters ψ) at the same step. The corresponding loss is defined as:

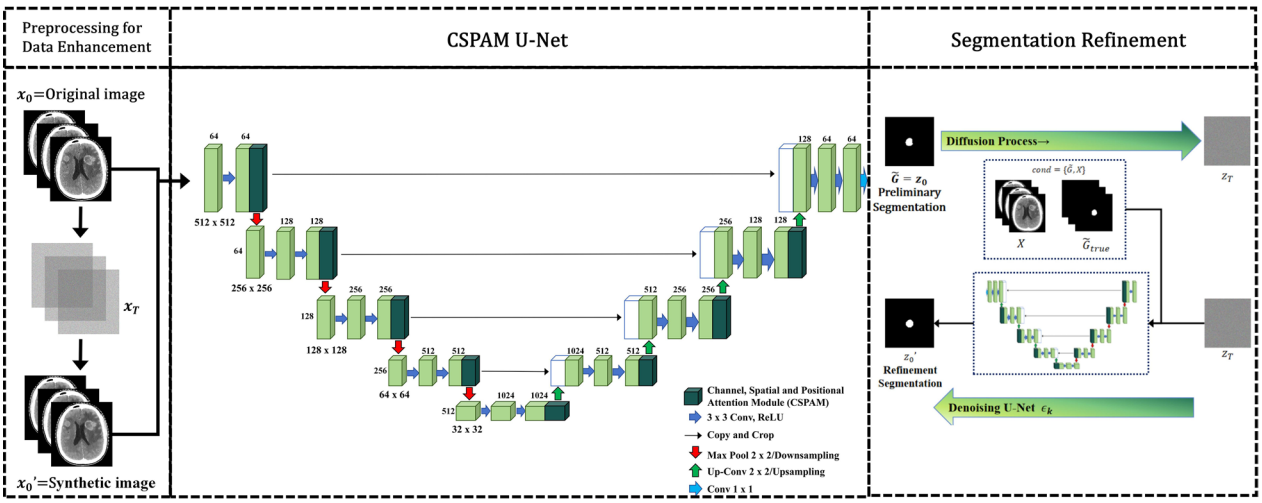

Fig. 3 Processing flow proposed in this study

$$L_{refine}\left(S_{refined}, S_{true}\right) = \mathbb{E}_{z_k,\varepsilon,cond}\varepsilon - \varepsilon_\psi(z_k, cond, k)^2$$

where z_k is the noisy segmentation mask at the k-th diffusion step, and $cond = \left\{\tilde{G}_{true}, X\right\}$ includes both the ground truth mask and the original image. Minimizing this objective function refines the initial segmentation produced by the U-Net. Combined with L_{seg}, the ultimate goal is to enhance the overall quality of the segmentation. The pipeline processing flow for this study is schematically shown as Fig. 3:

Model training and evaluation

All the model parameters, including the weights and biases of the convolutional layers, were initialized using the He method. The biases were initialized to 0. During each iteration, a batch of training data was fed into the model for forward propagation, and the output of the model was calculated. The model output and true labels were used to compute the loss function, which included the segmentation loss and boundary refinement loss. The gradients of the loss function were then calculated and backpropagated through the network to update the model parameters. Stochastic gradient descent (SGD) was used to optimize and update the network parameters, with the update formula for each parameter given by:

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta L(\theta),$$

where θ_t is the current parameter value, η is the learning rate, and $\nabla_\theta L(\theta)$ is the gradient of the loss function with respect to θ.

These steps were repeated over multiple training epochs until either the model converged or a predetermined number of iterations was reached. The maximum number of training epochs was set to 1000, the initial learning rate was 0.001, and the learning rate decay was 10^{-4}. The batch size was set to 8. The dataset from Jiangxi Cancer Hospital ($n=205$) was used for model training. fivefold cross-validation was used on the training set to internally validate the model performance. External validation was performed by using the dataset from the Affiliated Hospital of Southwest Medical University ($n=45$) to validate the generalization of the model. In the model evaluation of this study, we first compared the overall segmentation performance of the model, and the segmentation performance of different sizes of brain metastases. Subsequently, comparisons are made with the baseline U-Net model, the U-Net model with the addition of the CSPAM module proposed in this study, and the U-Net model with the addition of both the Diffusion Model and the CSPAM module. Finally, this study also includes four mainstream U-Net variant models and a Mask R-CNN model for comparing the model performance with our proposed model, including the improved U-Net model using the Attention Mechanism Module, Squeeze Excitation Module, the Residual Module, and the Transformer [20, 21, 26, 27, 47]. The performance of the models is evaluated using metrics such as Dice coefficient, intersection-to-union ratio (IoU), accuracy, sensitivity, specificity, and Hausdorff distance (HD). These metrics clearly indicate the performance of the model in the segmentation task.

Table 1 Numerical comparison among original images and synthetic images generated by different models

Method	PNSR (dB)	SSIM
Cycle GAN [40]	29.82 ± 6.25	0.84 ± 0.06
Pix2pix GAN [41]	30.54 ± 6.03	0.86 ± 0.10
Conditional GAN [42]	33.29 ± 5.16	0.90 ± 0.08
Ours	34.21 ± 4.70	0.91 ± 0.04

PSNR, Peak signal to noise ratio; SSIM, Structural similarity index

Statistical analysis

All performance metrics (DSC, IoU, Accuracy, Sensitivity, Specificity, and HD) were summarized as mean ± standard deviation. We compared each proposed model results (CSPAM-U-Net and Diffusion-CSPAM-U-Net) with the baseline U-Net results in the external validation dataset using the Wilcoxon signed-rank test. All statistical analyses were performed using Python (version 3.8) with the SciPy library (version 1.4.1).

Results

Quantitative comparison of synthetic CT image

In this study, we compared the proposed method with several existing generative adversarial network (GAN) models in terms of image synthesis quality. The quantitative results are presented in Table 1. We employed peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) as metrics to evaluate the overall image quality and structural fidelity. As shown in the table, the average PSNR of images synthesized by Cycle GAN [40] and Pix2pix GAN [41] reached 29.82 ± 6.25 dB and 30.54 ± 6.03 dB, respectively, with SSIM values of 0.84 ± 0.06 and 0.86 ± 0.10. Conditional GAN [42] showed further improvement, achieving a PSNR of 33.29 ± 5.16 dB and an SSIM of 0.90 ± 0.08. In contrast, our method attained the highest performance, with a PSNR of 34.21 ± 4.70 dB and an SSIM of 0.91 ± 0.04, surpassing the competing methods in both noise robustness and structural fidelity. These results confirm the effectiveness and robustness of our approach in medical image synthesis applications.

Segmentation model complexity and computational performance

To evaluate the feasibility of our proposed attention-based framework in resource-constrained or time-sensitive scenarios, we compared the model complexity and computational performance across different network configurations, as shown in Table 2. The baseline U-Net comprises 31.75 million parameters (M) and requires approximately 31.29 GigaFLOPs (G) per forward pass; under our training setup, it achieves a total training time of 104.63 h while sustaining an inference speed of 20 ms per slice. By progressively incorporating attention modules, we observe incremental increases in parameter count and computational cost. Specifically, integrating the Channel Attention Module (CAM) (U-Net + CAM) slightly raises the parameter count from 31.75 M to 31.81 M, but the FLOPs jump from 31.29G to 47.67G, reflecting the higher arithmetic intensity of channel-wise feature recalibration. Similarly, adding the Spatial Attention Module (SAM) (U-Net + CAM + SAM) raises the FLOPs to 42.83G and extends the total training time to 107.80 h, with an inference latency of 29 ms per slice. When Positional Attention Module (PAM) is further introduced (CSPAM-U-Net), the model reaches 32.30 M parameters and 58.05G FLOPs, incurring a total training time of 110.17 h and a slice-level inference time of 35 ms. Despite this additional overhead, the advanced attention synergy (channel, spatial, and positional) potentially leads to more robust feature representations and improved segmentation accuracy.

Overall model performance evaluation

The proposed diffusion-CSPAM-U-Net model demonstrated superior performance in both internal and external validations compared with the baseline U-Net and CSPAM-U-Net models (Table 3). In internal validation,

Table 2 Comparison of model complexity and computational performance

Model	Params (M)	FLOPs (G)	Total Training Time (h)	Inference Time/Per Slice (ms)
U-Net	31.75 M	31.29G	104.63 h	20 ms
U-Net + CAM	31.81 M	47.67G	105.52 h	27 ms
U-Net + CAM + SAM	31.94 M	42.83G	107.80 h	29 ms
U-Net + CAM + SAM + PAM (CSPAM-U-Net)	32.30 M	58.05G	110.17 h	35 ms

Note: The total training time included the diffusion model. The inference time is the time of inference per slice on the validation set for the completed trained model

CAM, Channel Attention Module; SAM, Spatial Attention Module; PAM, Positional Attention Module; Params, Parameters; GFLOPs, Giga Floating-point Operations Per Second

Table 3 Overall mean results of automated segmentation of brain metastases

	DSC	IoU	Accuracy	Sensitivity	Specificity	HD (mm)
Internal validation						
U-Net	0.760 ± 0.138	0.652 ± 0.130	0.907 ± 0.114	0.745 ± 0.129	0.894 ± 0.142	7.526 ± 1.107
CSPAM-U-Net	0.801 ± 0.126	0.700 ± 0.132	0.951 ± 0.110	0.792 ± 0.111	0.940 ± 0.145	6.430 ± 0.986
Diffusion-CSPAM-U-Net	0.844 ± 0.128	0.731 ± 0.125	0.972 ± 0.096	0.838 ± 0.113	0.972 ± 0.138	5.107 ± 0.984
External validation						
U-Net	0.698 ± 0.151		0.886 ± 0.111	0.708 ± 0.124	0.821 ± 0.147	8.324 ± 1.225
CSPAM-U-Net	0.756 ± 0.139[a]	0.656 ± 0.115	0.937 ± 0.107	0.761 ± 0.126	0.890 ± 0.139[a]	6.819 ± 1.104[a]
Diffusion-CSPAM-U-Net	0.793 ± 0.133[b]	0.692 ± 0.133[b]	0.955 ± 0.118	0.803 ± 0.121[b]	0.938 ± 0.140[b]	5.606 ± 0.990[b]

DSC, Dice similarity coefficient; IoU, intersection over union. "*a*" represents the comparison with U-Net where CSPAM-U-Net is significantly different in the corresponding metrics ($P < 0.05$). "*b*" denotes that Diffusion-CSPAM-U-Net is significantly different in the comparison with U-Net on the corresponding metric ($P < 0.05$)

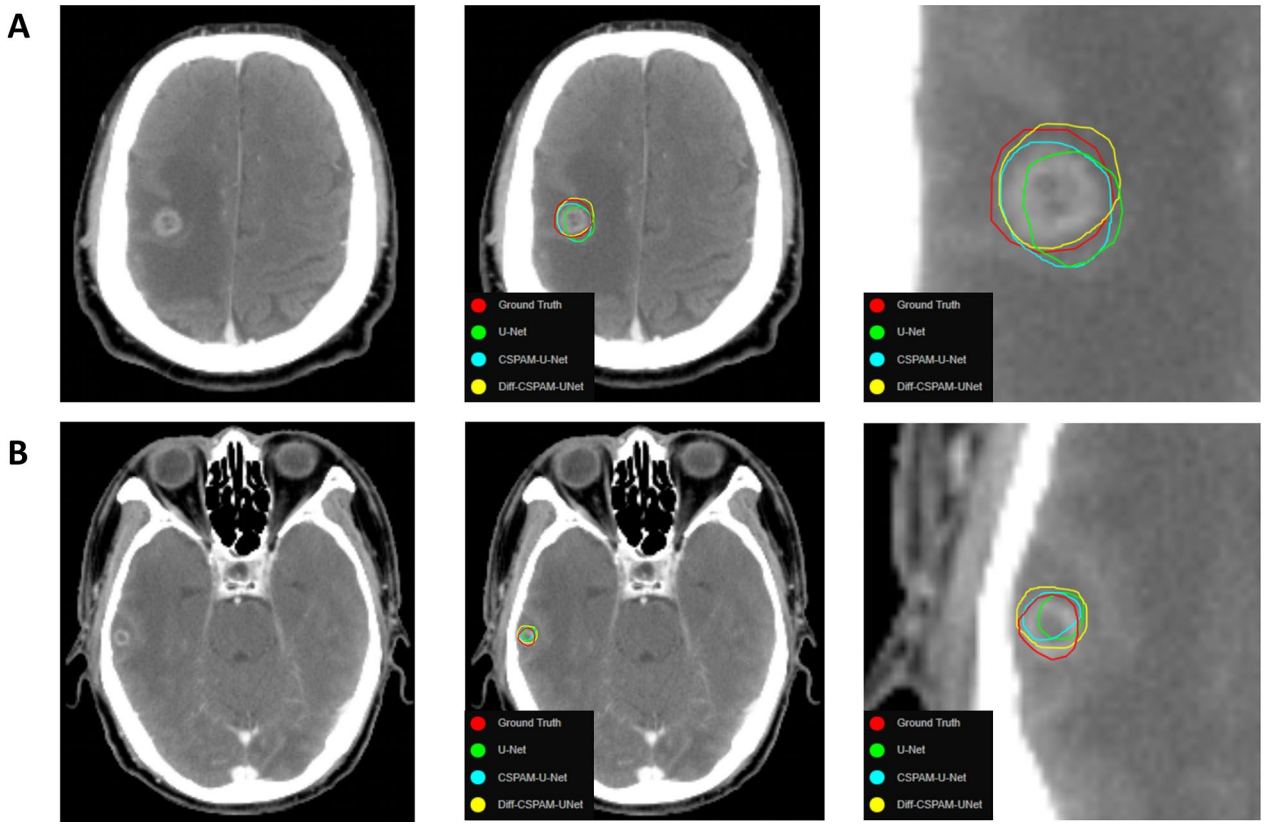

Fig. 4 Random samples were selected to demonstrate the segmentation results of the model proposed in this study. **A** From the external validation dataset. **B** From the training dataset

the diffusion-CSPAM-U-Net achieved a Dice similarity coefficient (DSC) of 84.4% ± 12.8%, intersection over union (IoU) of 73.1% ± 12.5%, accuracy of 97.2% ± 9.6%, sensitivity of 83.8% ± 11.3%, specificity of 97.2% ± 13.8%, and Hausdorff distance (HD) of 5.107 ± 0.984 mm. These results indicate a substantial improvement in capturing the true extent of brain metastases and minimizing segmentation errors. In the external validation, the model maintained robust performance with a DSC of 79.3% ± 13.3%, IoU of 69.2% ± 13.3%, accuracy of 95.5% ± 11.8%, sensitivity of 80.3% ± 12.1%, specificity of 93.8% ± 14.0%, and HD of 5.606 ± 0.990 mm. These

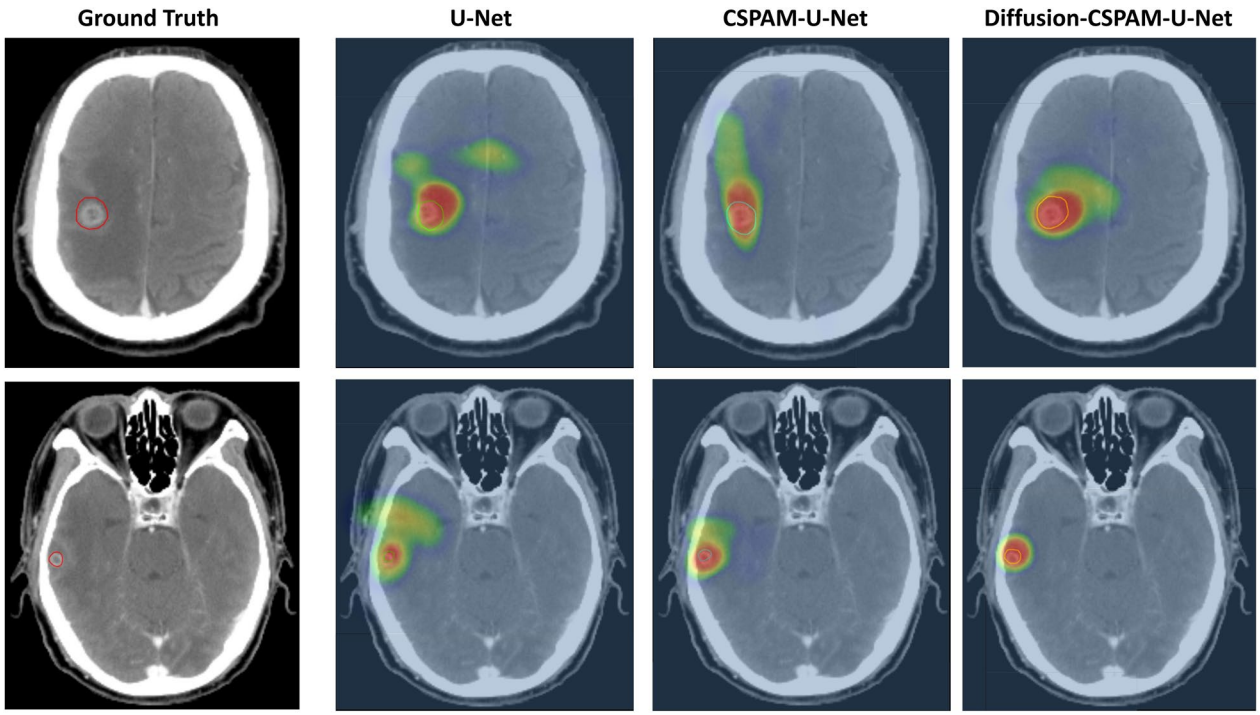

Fig. 5 Gradient-weighted class activation mapping (Grad-CAM)

Table 4 Comparison of segmentation results between the proposed model and those of other studies for brain metastases of different sizes in the external validation dataset

	DSC	IoU	Accuracy	Sensitivity	Specificity
Proposed model					
< 5 mm	0.652 ± 0.163	0.549 ± 0.158	0.780 ± 0.136	0.657 ± 0.158	0.780 ± 0.133
5–20 mm	0.775 ± 0.141	0.669 ± 0.144	0.936 ± 0.120	0.794 ± 0.127	0.924 ± 0.128
> 20 mm	0.816 ± 0.118	0.710 ± 0.140	0.969 ± 0.113	0.871 ± 0.115	0.961 ± 0.124
PAM-U-Net [21]					
< 5 mm	0.624 ± 0.168	0.526 ± 0.146	0.805 ± 0.123	0.630 ± 0.114	0.732 ± 0.139
5-20 mm	0.726 ± 0.121	0.630 ± 0.137	0.927 ± 0.115	0.712 ± 0.106	0.909 ± 0.116
> 20 mm	0.753 ± 0.094	0.698 ± 0.119	0.963 ± 0.106	0.747 ± 0.098	0.974 ± 0.103
GAN + Mask-R CNN + CRF [20]					
< 5 mm	0.635 ± 0.090	0.512 ± 0.058	0.776 ± 0.071	0.618 ± 0.102	0.839 ± 0.095
5-20 mm	0.710 ± 0.121	0.624 ± 0.104	0.875 ± 0.101	0.731 ± 0.105	0.912 ± 0.112
> 20 mm	0.749 ± 0.098	0.675 ± 0.082	0.938 ± 0.113	0.850 ± 0.089	0.967 ± 0.107

DSC, Dice similarity coefficient; IoU, intersection over union

metrics confirm the ability of the model to generalize well to new datasets, and demonstrate its generalizability to real clinical settings.

The visualization of the segmentation effect of the model proposed in this study is shown in Fig. 4. To further elucidate the performance and interpretability of the proposed diffusion-CSPAM-U-Net model, gradient-weighted class activation mapping (Grad-CAM) was employed to visualize the focus areas during segmentation. Figure 5 provides a comparative view of the Grad-CAM results for the diffusion-CSPAM-U-Net, CSPAM-U-Net, and U-Net models on randomly selected images from the datasets. The Grad-CAM visualizations show that the diffusion-CSPAM-U-Net model exhibits a more concentrated and accurate focus on the tumor regions than do the other models. This

enhanced focus contributes to higher Dice similarity coefficient (DSC) and intersection over union (IoU) scores, indicating the superior ability of the model to accurately delineate tumor boundaries.

Subgroup model performance results based on different sizes of brain metastases

This study evaluated the proposed model on brain metastases of different sizes revealed its effectiveness under various conditions (Table 4). In this study, the tumor size was determined by evaluating each metastasis on a slice-by-slice basis to identify the cross-sectional slice exhibiting the greatest extent of the lesion. On this identified slice, the maximum in-plane diameter was measured, reflecting the largest linear distance across the tumor. For metastases smaller than 5 mm, the model achieved a DSC of 65.2% ± 16.3%, IoU of 54.9% ± 15.8%, accuracy of 78.0% ± 13.6%, sensitivity of 65.7% ± 15.8%, and specificity of 78.0% ± 13.3%. These results indicate that although the model performs reasonably well on smaller lesions, there is room for improvement in accurately segmenting small metastases. For medium-sized metastases (5–20 mm), the model's performance improved significantly with a DSC of 77.5% ± 14.1%, IoU of 66.9% ± 14.4%, accuracy of 93.6% ± 12.0%, sensitivity of 79.4% ± 12.7%, and specificity of 92.4% ± 12.8%. This suggests that the model can effectively handle intermediate-sized lesions with improved accuracy and precision. For larger metastases (> 20 mm), the model achieved the highest performance with a DSC of 81.6% ± 11.8%, IoU of 71.0% ± 14.0%, accuracy of 96.9% ± 11.3%, sensitivity of 87.1% ± 11.5%, and specificity of 96.1% ± 12.4%. The high metrics in this category demonstrate the robustness of the model in accurately delineating larger tumor volumes, which are often more critical in clinical decision-making.

Comparison of overall performance with other brain metastases computed tomography image segmentation models

The diffusion-CSPAM-U-Net model outperformed other state-of-the-art models in terms of several key metrics (Table 5). Compared with the GAN + Mask-R-CNN + CRF model, which had a DSC of 72.6% ± 12.8% and IoU of 64.0% ± 13.6%, the proposed model achieved higher DSC (79.3% ± 13.3%) and IoU (69.2% ± 13.3%) values. This improvement highlights the enhanced capability of the proposed model to accurately segment brain metastases. A comparison with PAM-U-Net further emphasizes the superiority of diffusion-CSPAM-U-Net. While the PAM-U-Net achieved a DSC of 75.3% ± 17.2% and an IoU of 67.2% ± 15.9%, the proposed model showed better performance in both metrics. Additionally, the proposed model demonstrated higher accuracy (95.5% ± 11.8% versus 94.8% ± 12.5%), better sensitivity (80.3% ± 12.1% versus 72.1% ± 11.6%), and competitive specificity (93.8% ± 14.0% versus 96.3% ± 10.4%). Moreover, the proposed model's HD of 5.606 ± 0.990 mm in external validation is notably lower than the HD values reported for other models, indicating a finer and more precise boundary delineation capability. This characteristic is critical for improving the clinical usability of the segmentation outputs, particularly in treatment

Table 5 Comparison of previous studies

	Proposed Model	GAN + Mask-R-CNN + CRF [20]	PAM-U-Net [21]	ST-U-Net [21]	SEA-U-Net [21]	SERR-U-Net [21]
DSC	0.793 ± 0.133	0.726 ± 0.128	0.753 ± 0.172	0.747 ± 0.158	0.730 ± 0.135	0.718 ± 0.156
IoU	0.692 ± 0.133	0.640 ± 0.136	0.672 ± 0.159	0.667 ± 0.143	0.648 ± 0.150	0.625 ± 0.141
Accuracy	0.955 ± 0.118	0.915 ± 0.118	0.948 ± 0.125	0.930 ± 0.131	0.919 ± 0.118	0.898 ± 0.122
Sensitivity	0.803 ± 0.121	0.765 ± 0.131	0.721 ± 0.116	0.749 ± 0.120	0.702 ± 0.131	0.694 ± 0.126
Specificity	0.938 ± 0.140	0.922 ± 0.117	0.963 ± 0.104	0.951 ± 0.112	0.978 ± 0.106	0.946 ± 0.114
HD	5.606 ± 0.990	7.356 ± 0.603	6.912 ± 0.620	7.241 ± 0.835	7.539 ± 0.547	7.706 ± 0.728

DSC, Dice similarity coefficient; IoU, intersection over union; HD, Hausdorff distance

Table 6 Comparison of different channel and spatial attention pooling methods

Methods	DSC (%)	IoU (%)	Params (M)	GFLOPs (G)
CSPAM-U-Net (AvgPool)	72.8%	61.3%	32.30 M	52.39G
CSPAM-U-Net (MaxPool)	71.2%	60.4%	32.30 M	56.87G
CSPAM-U-Net (AvgPool & MaxPool)	75.6%	65.6%	32.30 M	58.05G

DSC, Dice Coefficient; IoU, Intersection over Union; Params, Parameters; GFLOPs, Giga Floating-point Operations Per Second

planning and monitoring. Despite these strengths, the diffusion-CSPAM-U-Net model has some areas in which it does not outperform the other models. For instance, the model's specificity, while competitive, is lower than that of SEA-U-Net (97.8% ± 10.6%) and SERR-U-Net (94.6% ± 11.4%). Higher specificity is important for minimizing false positives, and the slightly lower specificity here suggests that the diffusion-CSPAM-U-Net model might produce more false positive segments than do these models.

Ablation experiments with pooling approaches for channels and spatial attention modules

To evaluate the impact of pooling strategies in channel and spatial attention modules, we conducted ablation experiments with three variants of CSPAM-U-Net: (1) using average pooling (AvgPool) only, (2) using max pooling (MaxPool) only, and (3) combining both pooling operations. The quantitative comparisons are summarized in Table 6. The experimental results demonstrate that the hybrid pooling strategy (AvgPool & MaxPool) achieves the highest segmentation performance, with a Dice Similarity Coefficient (DSC) of 75.6% and an Intersection-over-Union (IoU) of 65.6%. This represents a significant improvement of 2.8% in DSC and 4.3% in IoU compared to the AvgPool-only variant (72.8% DSC, 61.3% IoU), and an even larger margin over the MaxPool-only variant (71.2% DSC, 60.4% IoU). The performance gap suggests that combining complementary pooling operations enables more robust feature aggregation. Specifically, AvgPool preserves global contextual information while MaxPool emphasizes salient local features, leading to enhanced attention maps. Notably, all three variants maintain identical parameter counts (32.30 M), indicating that the performance gains stem from architectural improvements rather than increased model capacity. However, the hybrid pooling approach incurs a moderate computational cost increase (58.05G GFLOPs) compared to AvgPool-only (52.39G GFLOPs) and MaxPool-only (56.87G GFLOPs) configurations. This balance between accuracy and computational complexity suggests that the combined pooling strategy effectively utilizes multi-scale information for attention refinement, justifying its adoption in our final architecture.

Discussion

In this study, the proposed diffusion-CSPAM-U-Net model demonstrated promising results in the segmentation of brain metastases in CT images. This study aimed to aid radiation oncologists in resource-limited areas of developing countries where MRI-sim is not available in GTV segmentation. The proposed diffusion-CSPAM-U-Net model, compared with CSPAM-U-Net in external validation, improved the DSC, IoU, sensitivity, specificity, and HD metrics by 4.9%, 5.5%, 5.5%, and 5.4%, respectively, and it reduced the HD by 1.213 mm. Compared with U-Net, the improvements in the DSC, IoU, sensitivity, specificity, and HD values were 13.6%, 17.9%, 13.4%, 14.2%, and 2.718 mm, respectively. These results indicate significant improvements in both capturing the true extent of brain metastases and reducing segmentation errors, thereby demonstrating their great potential for clinical applications. Notably, the diffusion-CSPAM-U-Net model significantly improved sensitivity to 80.3% ± 12.1%, showcasing the effectiveness of integrating diffusion models and channel-spatial-positional attention modules. This means that the model is more sensitive in detecting brain metastases, which thereby reduces the risk of missing GTV segmentation.

Compared with the results of previous studies, the diffusion-CSPAM-U-Net model performed excellently in multiple key metrics. Compared with the GAN + Mask-R-CNN + CRF model proposed by Wang et al. [20], the diffusion-CSPAM-U-Net model improved the Dice similarity coefficient (DSC) by 8.2%, intersection over union (IoU) by 8.1%, accuracy by 4.4%, sensitivity by 5.0%, and specificity by 1.7%. Compared with PAM-U-Net, the DSC improved by 5.3%, IoU by 3.0%, accuracy by 0.7%, and sensitivity by 8.2%. Although the specificity slightly decreased, it remained at a high level (93.8% versus 96.3%). The ResNet-101 backbone architecture of the GAN + Mask-R-CNN + CRF model was built through pretraining. Although pretraining improves the initial performance of the model to an extent, it also has some limitations. First, pretrained models are typically trained on large datasets (such as ImageNet), which have significant differences in feature distribution and imaging patterns compared with medical image datasets [22]. Therefore, pretrained models may not fully capture specific features and details in medical images when transferred to medical image segmentation tasks, which thereby limits segmentation performance [23]. Additionally, the GAN + Mask-R-CNN + CRF model relies on a multistage processing pipeline (i.e., GAN generates samples, Mask-R-CNN performs segmentation, and CRF performs post-processing), which can lead to error accumulation, with errors at each stage affecting the final segmentation result [24]. The training method used by diffusion-CSPAM-U-Net integrates image enhancement and segmentation into a unified framework that reduces the risk of error propagation and improves the overall segmentation performance.

Compared with the models built on the transformer architecture (ST-U-Net) and those improved with the squeeze-and-excitation (SE) module (SEA-U-Net and SERR-U-Net), diffusion-CSPAM-U-Net exhibited

superior performance in terms of DSC, IoU, accuracy, and sensitivity [25–27]. In particular, diffusion-CSPAM-U-Net significantly improved from 10.1 to 15.4%, which is crucial for reducing the risk of missed diagnoses. Although the specificity was slightly lower than those of SEA-U-Net and SERR-U-Net, the overall performance of diffusion-CSPAM-U-Net was better, especially in terms of segmentation accuracy and sensitivity. The ST-U-Net model, which is based on the transformer architecture, performs well in capturing global contextual information; however, transformer architectures typically require large-scale sample data for training to fully realize their potential [28]. This is a challenge for medical image datasets because obtaining high-quality annotated data is often time-consuming and costly. Therefore, the performance of ST-U-Net on small-sample datasets may be limited, whereas diffusion-CSPAM-U-Net adapts better to feature extraction and segmentation tasks under small-sample conditions by introducing diffusion models and attention mechanisms. The SEA-U-Net and SERR-U-Net models use a squeeze-and-excitation (SE) module to enhance feature representation capabilities. The SE module adapts by recalibrating the weights of the feature channels to thereby enhance the focus of the model on important features [29, 30]. However, the introduction of the SE module also increases the complexity and computational load of the model, which may reduce the efficiency of high-resolution image processing. Additionally, although the SE module enhances the selectivity of the feature channels, its effect may be limited when dealing with CT images with complex backgrounds and high noise levels [31].

In the comparison of segmentation performance across different sizes of brain metastases, the diffusion-CSPAM-U-Net model performed well in each subgroup but also showed some advantages and disadvantages. For metastases smaller than 5 mm, the diffusion-CSPAM-U-Net's Dice similarity coefficient (DSC) was 0.652 ± 0.163, which was significantly higher than Wang et al.'s PAM-U-Net (0.624 ± 0.168) and GAN + Mask-R-CNN + CRF (0.635 ± 0.090). This indicates that diffusion-CSPAM-U-Net can better capture tumor boundaries when dealing with small lesions to thereby improve the detection sensitivity and accuracy. However, its specificity was 0.780 ± 0.133, which was slightly lower than GAN + Mask-R-CNN + CRF's 0.839 ± 0.095, indicating room for improvement in reducing false positives. Nevertheless, the diffusion-CSPAM-U-Net model still had an advantage in terms of sensitivity (0.657 ± 0.158), showing better performance in detecting small lesions. Overall, diffusion-CSPAM-U-Net demonstrated significant advantages in the segmentation of brain metastases of different sizes, particularly in terms of sensitivity and segmentation accuracy. Its excellent performance for metastases smaller than 5 mm and 5–20 mm in size is particularly noteworthy because detecting these lesions is challenging and requires a higher discriminative ability from the model.

Despite the outstanding performance of the diffusion-CSPAM-U-Net model in multiple aspects, it still has a lower specificity than do models such as ST-U-Net based on the transformer architecture, SEA-U-Net, SERR-U-Net using squeeze-and-excitation mechanisms, and PAM-U-Net using individual positional attention modules. The noise and diversity introduced by the diffusion model during the preprocessing stage enhance image details but may also increase responses to non-tumor areas and thereby affect specificity [32]. The diffusion model improves image quality and model sensitivity by gradually adding noise and learning to reconstruct images from noise. However, when processing low-quality CT images, the introduced noise may cause normal tissues to be misidentified as tumors. Specifically, the diffusion model may enhance background noise and non-tumor structural features while enhancing image details, thereby increasing the probability of false positives. Although this method significantly improves the sensitivity of detecting small lesions and lesions with blurred edges, it also has a negative impact on the specificity of the model. Additionally, the use of channel-spatial-positional attention modules (CSPAM) enhances the model's ability to capture important features and spatial information, but it may also increase sensitivity to background noise. The CSPAM modules enhance the model's focus on specific features via channel attention, spatial attention, and positional attention mechanisms. These attention mechanisms can effectively focus on tumor areas and improve both segmentation accuracy and sensitivity [33]. However, in complex conditions of brain metastases, which often involve multiple lesions and complex brain tissue backgrounds, attention mechanisms may mistakenly identify certain background features as tumor features and thereby increase the false positive rate. For example, in images with complex backgrounds or high noise levels, attention mechanisms may overemphasize certain features of non-tumor areas and lead to misjudgment by the model.

Despite the promising performance of the diffusion-CSPAM-U-Net model in brain metastasis CT image segmentation, some limitations remain. First, the overall segmentation performance for small brain metastases (< 5 mm) remains limited. Additionally, as this is a dual-center study, and although it has achieved good results on an independent external validation set, the model may be affected when handling data from different devices, imaging parameters, and populations. Therefore,

cross-national multicenter studies are needed to further validate the generalizability of this model. Future research will optimize the parameters of the diffusion model and attention mechanisms to reduce the impact of background noise and improve specificity. The introduction of more intelligent noise processing methods and multilayer attention mechanisms should be considered to improve the ability of the model to distinguish between tumors and normal tissues. Additionally, additional brain metastasis CT image data from different sources, qualities, and populations should be collected internationally to enhance the generalizability and robustness of the model.

Conclusion

The proposed diffusion-CSPAM-U-Net model demonstrates significant improvements in the segmentation of brain metastases in CT images, especially in resource-limited environments where MRI-sim is not available, and it provides a helpful tool for radiation oncologists performing segmentation. By integrating diffusion and channel-spatial-positional attention mechanisms, the model significantly improves performance metrics compared with those of existing models.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13014-025-02622-x.

Supplementary material 1.

Acknowledgements
Not applicable.

Author contributions
(I) Conception and design: Haowen Pang, Ping Zhou, Huaiwen Zhang, Yiren Wang; (II) Administrative support: Haowen Pang, Ping Zhou, Huaiwen Zhang; (III) Provision of study materials or patients: Haowen Pang, Huaiwen Zhang; (IV) Collection and assembly of data: Yiren Wang, Zhongjian Wen, Shuilan Bao, Delong Huang, Youhua Wang, Bo Yang and Yunfei Li; (V) Data analysis and interpretation: Yiren Wang, Zhongjian Wen, Shuilan Bao and Delong Huang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

Funding
Xuyong County People's Hospital-Southwest Medical University Science and Technology Strategic Cooperation Program (No.2024XYXNYD05), Gulin County People's Hospital-Affiliated Hospital of Southwest Medical University Science and Technology Strategic Cooperation Program (No.2022GLXNYDFY05), Key-funded Project of the National College Student Innovation and Entrepreneurship Training Program (No.202310632001), National College Student Innovation and Entrepreneurship Training Program (No.S202410632165X), National College Student Innovation and Entrepreneurship Training Program (No.2024391), Sichuan Medical and Health CarePromotion Institute Project (project number: KY2022SJ0377), Luzhou-Southwest Medical University Transformation and Landing Support Project (No. 2024LZXNYDZ002). The Open Fund for Scientific Research of Jiangxi Cancer Hospital (No.2021J15).

Availability of data and materials
The data used and/or analyzed during the current study available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declare no competing interests.

Author details
[1]School of Nursing, Southwest Medical University, Luzhou 646000, China. [2]Wound Healing Basic Research and Clinical Application Key Laboratory of Luzhou, School of Nursing, Southwest Medical Unversity, Luzhou 646000, China. [3]Department of Clinical Medicine, Southwest Medical University, Luzhou 646000, China. [4]Gulin County People's Hospital, Luzhou 646000, China. [5]Department of Oncology, The Affiliated Hospital of Southwest Medical University, No.25 Taiping Street, Jiangyang District, Luzhou 646000, Sichuan, China. [6]Department of Radiology, The Affiliated Hospital of Southwest Medical University, No.25 Taiping Street, Jiangyang District, Luzhou 646000, Sichuan, China. [7]Department of Radiotherapy, Jiangxi Cancer Hospital, The Second Affiliated Hospital of Nanchang Medical College, Jiangxi Clinical Research Center for Cancer, No.519 Beijing East Road, Donghu District, Nanchang 330029, Jiangxi, China.

Received: 15 June 2024 Accepted: 11 March 2025
Published online: 05 April 2025

References

1. Boire A, Brastianos PK, Garzia L, Valiente M. Brain metastasis. Nat Rev Cancer. 2020;20(1):4–11.
2. Lamba N, Wen PY, Aizer AA. Epidemiology of brain metastases and leptomeningeal disease. Neuro Oncol. 2021;23(9):1447–56.
3. Suh JH, Kotecha R, Chao ST, Ahluwalia MS, Sahgal A, Chang EL. Current approaches to the management of brain metastases. Nat Rev Clin Oncol. 2020;17(5):279–99.
4. Moss NS, Beal K, Tabar V. Brain metastasis: a distinct oncologic disease best served by an integrated multidisciplinary team approach. JAMA Oncol. 2022;8(9):1252–4.
5. Castellano A, Bailo M, Cicone F, Carideo L, Quartuccio N, Mortini P, Falini A, Cascini GL, Minniti G. Advanced imaging techniques for radiotherapy planning of gliomas. Cancers. 2021;13(5):1063.
6. Pham J, Neilsen BK, Liu H, Cao M, Yang Y, Sheng K, Ma TM, Kishan AU, Ruan D. Dosimetric predictors for genitourinary toxicity in MR-guided stereotactic body radiation therapy (SBRT): substructure with fraction-wise analysis. Med Phys. 2024;51(1):612–21.
7. Martin CJ, Kron T, Vassileva J, Wood TJ, Joyce C, Ung NM, Small W, Gros S, Roussakis Y, Plazas MC, Benali AH. An international survey of imaging practices in radiotherapy. Physica Med. 2021;1(90):53–65.
8. McGee KP, Cao M, Das IJ, Yu V, Witte RJ, Kishan AU, Valle LF, Wiesinger F, De-Colle C, Cao Y, Breen WG. The use of magnetic resonance imaging in radiation therapy treatment simulation and planning. J Magnetic Resonance Imaging. 2024. https://doi.org/10.1002/jmri.29246.
9. Rhee DJ, Jhingran A, Rigaud B, Netherton T, Cardenas CE, Zhang L, Vedam S, Kry S, Brock KK, Shaw W, O'Reilly F. Automatic contouring system for cervical cancer using convolutional neural networks. Med Phys. 2020;47(11):5648–58.
10. Zhai X, Kolesnikov A, Houlsby N, Beyer L. Scaling vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2022 pp. 12104–12113
11. Esmaeilzadeh P. Challenges and strategies for wide-scale artificial intelligence (AI) deployment in healthcare practices: a perspective for healthcare organizations. Artif Intell Med. 2024;1(151): 102861.
12. Ali O, Ali H, Shah SA, Shahzad A. Implementation of a modified U-Net for medical image segmentation on edge devices. IEEE Trans Circuits Syst II Express Briefs. 2022;69(11):4593–7.
13. Du G, Cao X, Liang J, Chen X, Zhan Y. Medical image segmentation based on U-Net: a review. J Imaging Sci Technol. 2020;64(2):020508–11.

14. Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-net and its variants for medical image segmentation: a review of theory and applications. Ieee Access. 2021;3(9):82031–57.
15. Bhunia AK, Khan S, Cholakkal H, Anwer RM, Laaksonen J, Shah M, Khan FS. Person image synthesis via denoising diffusion model. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2023 pp. 5968–5976
16. Fan Y, Liao H, Huang S, Luo Y, Fu H, Qi H. A survey of emerging applications of diffusion probabilistic models in mri. Meta-Radiology. 2024;9: 100082.
17. Wang X, He Z, Peng X. Artificial-intelligence-generated content with diffusion models: a literature review. Mathematics. 2024;12(7):977.
18. He X, Tan C, Han L, Liu B, Axel L, Li K, Metaxas DN. DMCVR: Morphology-Guided Diffusion Model for 3D Cardiac Volume Reconstruction. InInternational Conference on Medical Image Computing and Computer-Assisted Intervention 2023 pp. 132–142. Cham: Springer Nature Switzerland
19. Guo J, Jia N, Bai J. Transformer based on channel-spatial attention for accurate classification of scenes in remote sensing image. Sci Rep. 2022;12(1):15473.
20. Wang Y, Wen Z, Su L, Deng H, Gong J, Xiang H, He Y, Zhang H, Zhou P, Pang H. Improved brain metastases segmentation using generative adversarial network and conditional random field optimization mask R-CNN. Med Phys. 2024;51(9):5990–6001.
21. Wang Y, Hu Y, Chen S, Deng H, Wen Z, He Y, Zhang H, Zhou P, Pang H. Improved automatic segmentation of brain metastasis gross tumor volume in computed tomography images for radiotherapy: a position attention module for U-Net architecture. Quant Imaging Med Surg. 2024;14(7):4475.
22. Spolaôr N, Lee HD, Mendes AI, Nogueira CV, Parmezan AR, Takaki WS, Coy CS, Wu FC, Fonseca-Pinto R. Fine-tuning pre-trained neural networks for medical image classification in small clinical datasets. Multimed Tools Appl. 2024;83(9):27305–29.
23. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: a review of deep learning solutions for medical image segmentation. Med Image Anal. 2020;1(63): 101693.
24. Katakis S, Barotsis N, Kakotaritis A, Economou G, Panagiotopoulos E, Panayiotakis G. Automatic extraction of muscle parameters with attention UNet in ultrasonography. Sensors. 2022;22(14):5230.
25. He X, Zhou Y, Zhao J, Zhang D, Yao R, Xue Y. Swin transformer embedding UNet for remote sensing image semantic segmentation. IEEE Trans Geosci Remote Sens. 2022;19(60):1–5.
26. Xiong L, Yi C, Xiong Q, Jiang S. SEA-NET: medical image segmentation network based on spiral squeeze-and-excitation and attention modules. BMC Med Imaging. 2024;24(1):17.
27. Wang J, Li X, Lv P, Shi C. SERR-U-Net: squeeze-and-excitation residual and recurrent block-based U-Net for automatic vessel segmentation in retinal image. Comput Math Methods Med. 2021. https://doi.org/10.1155/2021/5976097.
28. Gao Y, Zhou M, Liu D, Yan Z, Zhang S, Metaxas DN. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. arXiv preprint arXiv:2203.00131. 2022
29. Üzen H, Turkoglu M, Aslan M, Hanbay D. Depth-wise squeeze and excitation block-based efficient-Unet model for surface defect detection. Vis Comput. 2023;39(5):1745–64.
30. Rajendran T, Valsalan P, Amutharaj J, Jenifer M, Rinesh S, Anitha T. Hyperspectral image classification model using squeeze and excitation network with deep learning. Comput Intell Neurosci. 2022;2022:1–9.
31. Li J, Li H, Zhang Y, Wang Z, Zhu S, Li X, Hu K, Gao X. MCNet: a multi-level context-aware network for the segmentation of adrenal gland in CT images. Neural Netw. 2024;1(170):136–48.
32. Li H, Yang Y, Chang M, Chen S, Feng H, Xu Z, Li Q, Chen Y. Srdiff: single image super-resolution with diffusion probabilistic models. Neurocomputing. 2022;28(479):47–59.
33. Zhou X, Tang C, Huang P, Mercaldo F, Santone A, Shao Y. LPCANet: classification of laryngeal cancer histopathological images using a CNN with position attention and channel attention mechanisms. Interdis Sci Comput Life Sci. 2021;13(4):666–82.
34. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. Adv Neural Inf Process Syst. 2020;33:6840–51.
35. Nichol AQ, Dhariwal P. Improved denoising diffusion probabilistic models. In: International conference on machine learning 2021 pp. 8162–8171. PMLR.
36. Yuan X, Li L, Wang J, Yang Z, Lin K, Liu Z, Wang L. Spatial-Frequency U-Net for Denoising Diffusion Probabilistic Models. arXiv preprint arXiv:2307.14648. 2023
37. Khader F, Müller-Franzes G, Tayebi Arasteh S, Han T, Haarburger C, Schulze-Hagen M, Schad P, Engelhardt S, Baeßler B, Foersch S, Stegmaier J. Denoising diffusion probabilistic models for 3D medical image generation. Sci Rep. 2023;13(1):7303.
38. Peng J, Qiu RL, Wynne JF, Chang CW, Pan S, Wang T, Roper J, Liu T, Patel PR, Yu DS, Yang X. CBCT-Based synthetic CT image generation using conditional denoising diffusion probabilistic model. Med Phys. 2024;51(3):1847–59.
39. Chen X, Qiu RL, Peng J, Shelton JW, Chang CW, Yang X, Kesarwala AH. CBCT-based synthetic CT image generation using a diffusion model for CBCT-Guided lung radiotherapy. Med Phys. 2024. https://doi.org/10.1002/mp.17328.
40. Liang X, Chen L, Nguyen D, Zhou Z, Gu X, Yang M, Wang J, Jiang S. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. Phys Med Biol. 2019;64(12): 125002.
41. Aljohani A, Alharbe N. Generating synthetic images for healthcare with novel deep pix2pix gan. Electronics. 2022;11(21):3470.
42. Dar SU, Yurt M, Karacan L, Erdem A, Erdem E, Cukur T. Image synthesis in multi-contrast MRI with conditional generative adversarial networks. IEEE Trans Med Imaging. 2019;38(10):2375–88.
43. Woo S, Park J, Lee JY, Kweon IS. Cbam: Convolutional block attention module. In: Proceedings of the European conference on computer vision (ECCV) 2018 pp. 3–19
44. Fu J, Liu J, Tian H, Li Y, Bao Y, Fang Z, Lu H. Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition 2019 pp. 3146–3154
45. Zhu X, Cheng D, Zhang Z, Lin S, Dai J. An empirical study of spatial attention mechanisms in deep networks. In: Proceedings of the IEEE/CVF international conference on computer vision 2019 pp. 6688–6697
46. Luo J, Wang Q, Zou R, Wang Y, Liu F, Zheng H, Du S, Yuan C. A heart image segmentation method based on position attention mechanism and inverted pyramid. Sensors. 2023;23(23):9366.
47. Zhang J, Qin Q, Ye Q, Ruan T. ST-unet: swin transformer boosted U-net with cross-layer feature enhancement for medical image segmentation. Comput Biol Med. 2023;1(153): 106516.
48. 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 2022 pp. 10684–10695

Publisher's Note