SIN-CHN SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(3), 1431.
https://doi.org/10.62617/mcb1431

Article

CALF-GAN: Multi-scale convolutional attention for latent feature-guided cross-modality MR image synthesis

Xinmiao Zhu[1,*], Yuan Wang[2]

[1] School of Computer Science and Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China
[2] The First Affiliated Hospital of Ningbo University, Ningbo 315211, China
*** Corresponding author:** Xinmiao Zhu, 202220601015@mails.zstu.edu.cn

CITATION

Zhu X, Wang Y. CALF-GAN: Multi-scale convolutional attention for latent feature-guided cross-modality MR image synthesis. Molecular & Cellular Biomechanics. 2025; 22(3): 1431. https://doi.org/10.62617/mcb1431

ARTICLE INFO

Received: 23 January 2025
Accepted: 7 February 2025
Available online: 18 February 2025

Abstract: Multimodal medical image synthesis plays a crucial supportive role in research within the field of biomechanics, providing high-precision data and analytical methods for studies on anatomical structures, tissue characteristics, and mechanical modeling. However, due to practical constraints, certain modalities of medical images may be difficult to obtain, posing challenges for model training and high-accuracy biomechanical research. Existing methods employ convolutional neural network (CNN)-based generative adversarial models to synthesize missing modality information across modalities. However, CNNs are limited in their ability to model long-range dependencies. Transformers offer a new paradigm to address these limitations, yet their high computational and memory demands remain a significant drawback. To tackle these challenges, we propose a novel generative adversarial model, termed the Convolutional Attention Latent Feature GAN (CALF-GAN), which leverages multi-scale convolutional attention for cross-modal medical image synthesis. A dedicated latent attribute separation module is employed to disentangle modality-specific features between source and target modality images, enhancing the synthesis of medical semantics, such as pixel intensity values. Furthermore, to improve the model's capacity for long-range dependency modeling while reducing computational overhead, we design a generation module based on multi-scale convolutional attention, capturing long-range dependencies using only convolutional operations. Extensive experiments conducted on various medical image datasets demonstrate that CALF-GAN achieves remarkable generalizability and outstanding overall performance under low memory requirements, making it well-suited for application in high-precision biomechanics research.

Keywords: medical image synthesis; biomechanics; adversarial; generative; magnetic resonance imaging (MRI); latent space; attention

1. Introduction

With the continuous advancement of artificial intelligence algorithms, medical image recognition technology holds significant research value and application potential in the field of biomechanics [1]. By integrating modern medical image processing techniques with biomechanical analysis, a deeper understanding of the mechanical properties of human tissues and organs can be achieved, thereby advancing research and applications in biomechanics. Medical image recognition technology enables the extraction of precise geometric information of organs, tissues, and bones from imaging modalities such as Computed Tomography (CT,) MRI, and ultrasound, facilitating the generation of personalized three-dimensional models [2]. These models can be utilized to simulate the nonlinear mechanical behavior of complex organs, enhancing the accuracy of biomechanical simulations [3].

Furthermore, dynamic medical imaging (e.g., 4D ultrasound, functional MRI) combined with image recognition algorithms can capture the mechanical responses of tissues during motion or loading processes [4].

Multimodal medical imaging plays a pivotal role in biomechanical research, as well as in clinical diagnosis and treatment [5], and significantly contributes to the development of deep learning models supporting various biomechanical studies and medical tasks [6–9]. The integration of different imaging modalities, such as CT, MRI, and Positron Emission Tomography (PET), enables the provision of more comprehensive biomechanical information. Different medical imaging modalities offer distinct advantages. For example, multi-contrast magnetic resonance imaging (MRI) can effectively separate fat and water signals by adjusting contrast parameters, enabling the acquisition of more detailed soft tissue images. In spinal imaging, the contrast between fat and water signals allows clear identification of structural changes in intervertebral discs, the spinal cord, and surrounding tissues, which is particularly significant for evaluating degenerative spinal conditions. Similarly, in brain imaging, variations in fat and water signals enable more precise differentiation of brain tissue, tumors, and edema regions, providing critical insights for tumor staging and treatment planning. These features are essential for diagnosing and treating different pathological regions in modern medicine. However, due to patient-specific challenges, certain modality images are often difficult to acquire. For instance, issues such as patients being unable to cooperate, high acquisition costs, and privacy protocol requirements can prevent the collection of corresponding modality images [10,11]. Additionally, factors like patient positioning discomfort, equipment limitations, or motion artifacts may result in missing MRI images. Therefore, developing effective methods to obtain these missing modality images is of great significance.

To address this, cross-modal synthesis techniques [12–14] have emerged as a promising solution. In recent years, convolutional neural network (CNN)-based cross-modal synthesis methods [15,16] have significantly improved the quality of synthesized images for missing modalities, enabling their use in assisting diagnosis, treatment, and data augmentation for scarce samples. Simultaneously, generative adversarial networks (GANs) [17–19] have become a cornerstone in the field of cross-modal synthesis due to their exceptional versatility and realistic generation performance. Notably, CycleGAN [20], with its unique bidomain cycle consistency, has greatly enhanced the quality and stability of synthesized images, establishing itself as one of the most critical backbone networks for cross-modal medical image synthesis. However, the intrinsic local receptive field of CNNs imposes limitations on their ability to model long-range dependencies. The advent of Transformers [21] offers a novel paradigm to overcome these limitations. Many recent studies have explored the potential of Transformers [22–24] in the domain of cross-modal medical image synthesis, achieving state-of-the-art (SOTA) results. Despite their superior performance in modeling long-range dependencies, these methods come at the cost of substantial computational overhead.

To this end, numerous researchers have attempted to streamline attention mechanisms [25–27], aiming to achieve a favorable trade-off between long-range dependency modeling and computational complexity. These methods effectively provide high-quality medical image generation with enriched contextual and structural

semantic information. However, medical images often contain specific global semantic information, such as pixel intensity distributions across different modalities. The global nature of this semantic information makes it challenging for lightweight attention mechanisms to fully capture the overall characteristics of the images. Recently, attribute decomposition methods based on latent space [28–30] have offered a novel approach to addressing this issue. By projecting image information into a latent space, these methods enable the extraction and separation of both local and global attributes.

Building upon the aforementioned ideas, we propose a latent feature-guided generative adversarial model based on multi-scale convolutional attention. This framework consists of a guidance module leveraging latent modality feature information and a generation module built on multi-scale convolutional attention. Specifically, we introduce an attribute separation mechanism in the latent space to disentangle global semantic information across different modalities. This enhances the realism of cross-modal generation, particularly in terms of pixel intensity values. Additionally, we design a multi-scale convolutional attention module to achieve a favorable trade-off between low memory consumption and robust long-range dependency modeling. Extensive experiments conducted on various datasets demonstrate that the proposed model can generate high-resolution, high-fidelity target modality images with minimal computational cost, making it suitable for application in high-precision biomechanical research.

The main contributions of our study are as follows:

- We present a novel latent modality feature-guided approach for cross-modality medical image generation, which enables high-precision biomechanical analysis. By incorporating an attribute separation mechanism in the latent space, this method enhances the model's ability to capture global semantic information by leveraging the disentangled modality-specific features in the medical imaging domain.
- We introduce a multi-scale attention module based solely on convolutional operations, striking a favorable balance between generating high-fidelity target modality images and maintaining a lower computational complexity for the model.
- We design a feature encoding module that incorporates a multi-scale convolutional attention mechanism to enhance the quality of latent feature encoding and better preserve global semantic information.

2. Related work

The integration of cross-modality medical image synthesis with biomechanics introduces new opportunities for medical research, clinical diagnostics, and personalized treatment. By generating medical images across different modalities (such as CT, MRI, PET, and ultrasound) and incorporating biomechanical modeling and analysis, this approach significantly enhances the efficiency of image utilization and the precision of biomechanical studies [31].

2.1. GAN in cross-modal medical image synthesis

Cross-modal medical image synthesis is a highly promising application, enabling the prediction of missing modality information in the absence of target modality data. Unlike other generation methods, it leverages the structural and semantic information of existing modalities, which is crucial for ensuring the realism of medical images. In recent years, Generative Adversarial Networks (GANs) have rapidly emerged as a mainstream approach in medical imaging due to their remarkable ability to generate realistic images and continuously improving resolution capabilities. They have been widely applied in biomechanical research, such as the biomechanical analysis of muscle tissues during movement [32,33]. Notably, CycleGAN, with its innovative bidomain consistency model, has laid the foundation for cross-modal synthesis and has profoundly influenced most existing methods in cross-modal medical image generation [34–36]. Specifically, Yurt et al. [37] proposed a semi-supervised deep generative model based on GAN to synthesize high-quality MR images from undersampled data. Liang et al. [38] utilized CycleGAN to generate Cone-Beam Computed Tomography (CBCT) images from CT scans. Emami et al. [39] developed a structure-aware generative adversarial network (SA-GAN) to synthesize CT images from MRI data. However, GANs, which are predominantly built on CNN backbones, remain limited by the intrinsic inability of CNNs to model long-range dependencies effectively, leading to suboptimal performance in certain scenarios.

2.2. Transformer in cross-modal medical image synthesis

The emergence of Transformers has effectively addressed the limitations of modeling long-range dependencies. Specifically, Transformers leverage self-attention mechanisms to capture long-range dependencies by determining the relevance of all embedded patches. Recently, Vision Transformer (ViT) [40] introduced global self-attention mechanisms to the image domain, achieving notable success across various medical imaging tasks [41,42]. In particular, researchers have explored Transformers as a means to overcome GANs' shortcomings in modeling long-range dependencies in cross-modal medical image synthesis. Zhao et al. [24] proposed the Residual Transformer Conditional GAN (RTCGAN), which combines the strengths of both approaches to generate CT images from MR data. Similarly, Zhang et al. [43] introduced a novel MRI synthesis framework, the Pyramid Transformer Network (PTNet), to achieve this task. However, Transformers inherently introduce significant computational and memory overhead.

To mitigate these drawbacks, some researchers have begun exploring the integration of CNNs with attention mechanisms to achieve a balance between the superior local spatial capabilities of CNNs and the long-range dependency modeling power of Transformers, all under low computational loads. Li et al. [27] proposed an Efficient Spatial Reduction Attention (ESRA) mechanism to enhance feature extraction while reducing computational complexity. Xu et al. [26] further advanced this idea by designing the CFBlock, which utilizes learnable convolutional attention to extract contextual information from pre-trained Transformer blocks. Our approach similarly focuses on addressing the computational and memory challenges of Transformers by relying exclusively on convolutional operations, providing an

efficient alternative to maintain high performance in long-range dependency modeling.

CFBlock enhances contextual refinement in CNN-based architectures by using dilated convolutions to expand the receptive field and applying channel-wise and spatial attention to dynamically refine feature importance. Unlike our multi-scale convolutional attention mechanism, which explicitly integrates multi-scale information for comprehensive feature representation, CFBlock prioritizes context-aware refinement through dilation rather than multi-scale convolution.

Efficient Spatial Reduction Attention (ESRA) improves efficiency by using spatial downsampling and global pooling, employing a lightweight attention mechanism to reduce computational cost while retaining essential spatial information. Unlike the multi-scale convolutional attention mechanism we proposed, which enhances feature learning across multiple receptive fields while preserving spatial details, ESRA prioritizes low-cost efficiency by reducing spatial resolution.

2.3. Latent space in cross-modal medical image synthesis

In recent years, an increasing number of medical imaging studies have recognized the significant role of latent space in feature extraction, processing, and reducing computational costs. Chartsias et al. [15] proposed a fully convolutional neural network that embeds all input modalities into a shared, modality-invariant latent space for generation. Fetty et al. [44] explored the manipulation of latent space based on StyleGAN, demonstrating its potential for flexible control. Dalmaz et al. [12] utilized latent space for feature processing during the generation process, achieving a better balance between computational cost and generation performance. However, these methods face limitations in explicit attribute disentanglement. The facial synthesis domain has provided insightful paradigms for latent space attribute decomposition [28–30], which inspired our approach. Building upon these advancements, we propose a latent feature-guided generative adversarial model based on multi-scale convolutional attention. Specifically, we treat modality as a separable latent feature to enhance the global semantic information in cross-modal medical image synthesis. By incorporating features from the latent space, more precise biomechanical studies can be conducted.

3. Method

As shown in **Figure 1**, the proposed Convolutional Attention Latent Feature-Guided Generative Adversarial Network (CALF-GAN) for biomechanical research consists of two key components: a guidance module based on latent modality feature information (Section 3.1) and a multi-scale convolutional attention module (Section 3.2) incorporated into the generator architecture (Section 3.3). Furthermore, the discriminator and model loss functions will be detailed in Section 3.4.

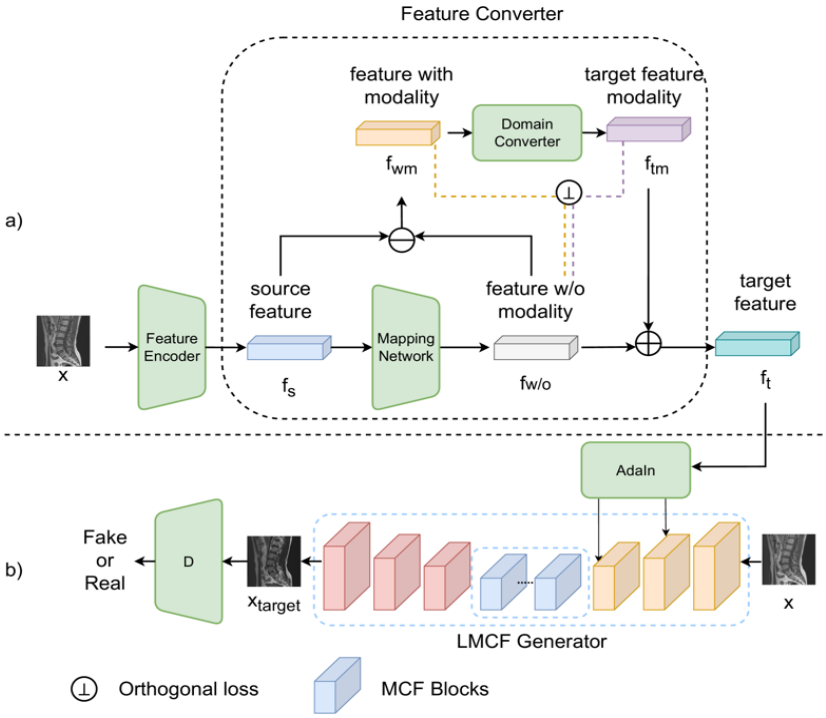

Figure 1. The proposed CALF-GAN architecture is built around two core components: **(a)** a guidance module based on latent modality feature information; **(b)** a generation module utilizing multi-scale convolutional attention.

3.1. The guidance module based on latent modality feature information

The proposed Guidance Module based on Latent Modality Feature Information is illustrated in **Figure 1a**. It consists of a Feature Encoder (FE) and a Feature Converter (FC). Specifically, the source modality image is encoded into a feature vector by the FE, which is then transformed into target modality feature information via the FC. The design of the FC is inspired by the Style Transformer in L2M-GAN [30], but we have removed redundant multi-attribute components, focusing instead on factorizing only the modality-specific information of the source and target images in the latent feature space. This modality-specific information is then used to guide the generator in cross-modality synthesis. To ensure the separation of modality feature vectors from unrelated feature vectors, we also employ an orthogonal loss.

Orthogonal loss explicitly enforces feature independence by penalizing similarity, ensuring diverse and meaningful representations. Compared to other disentanglement strategies like adversarial disentanglement and mutual information minimization, orthogonal loss is more efficient and easier to integrate into standard training frameworks. It enhances generalization and robustness by reducing feature redundancy, improving interpretability, and maintaining computational stability without requiring additional network components. These advantages make it a practical and effective choice for feature separation approach.

The feature vector f extracted from the source modality image is crucial for the quality of subsequent transformations. In particular, preserving the structural semantic information of medical images is vital for the overall generation process. Therefore, we have designed an FE block that incorporates a multi-scale convolutional attention mechanism, as shown in **Figure 2**. Specifically, we add our proposed Multi-Conv-Former (MCF) module to the feature extraction layer, after eliminating redundant

residual blocks. This module employs a multi-scale convolutional attention mechanism to extract semantic information from medical images using only convolution operations. This approach provides better integration support for separating structural semantic information and modality-specific information in the FC module.

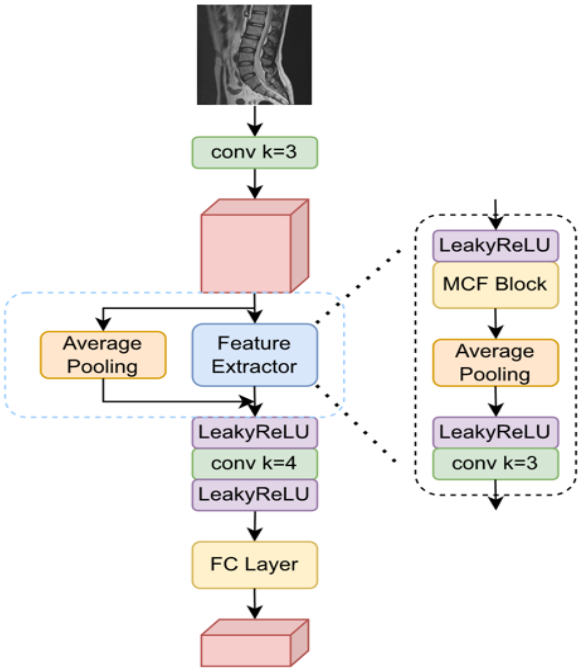

Figure 2. The feature encoding (FE) module incorporating a multi-scale convolutional attention mechanism.

3.2. Multi-scale convolution attention

To better simulate the long-range dependency modeling capability of Transformers, we have designed the Multi-Conv-Former module, as shown in **Figure 3b**.

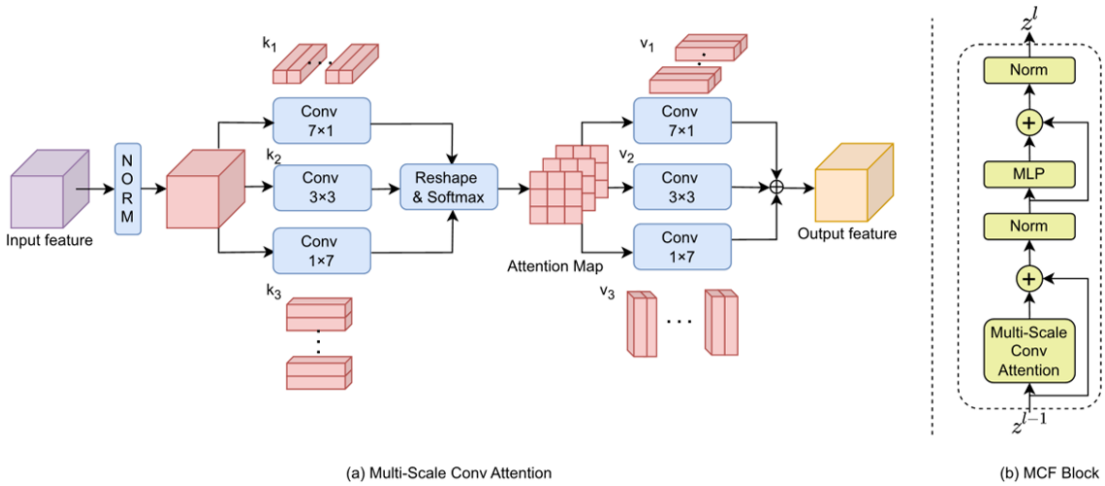

Figure 3. (a) Architecture diagram of the multi-scale convolutional attention (MSCA) module; **(b)** architecture diagram of the multi-conv-former block.

Specifically, we adopt the structural principles of the Transformer but replace the self-attention module with the multi-scale convolutional attention mechanism that we propose. This allows for more efficient learning of the semantic information of medical images with lower memory usage and computational cost. The corresponding formula is as follows:

$$\widetilde{z^l} = \mathrm{BN}\left(\mathrm{MSCA}\left(z^{l-1}\right) + z^{l-1}\right) \tag{1}$$

$$z^l = \mathrm{BN}\left(\mathrm{MLP}\left(\widetilde{z^l}\right) + \widetilde{z^l}\right) \tag{2}$$

here, MSCA refers to the Multi-Scale Convolution Attention module, as shown in **Figure 3a**. BN represents batch normalization [45], z^{l-1} and z^l denote the input and output, and $\widetilde{z^l}$ represents the hidden features output by the MSCA.

The structural semantic information in medical images plays a critical role in the quality of generation. Our approach is inspired by Xu et al.'s [26] attempt to extract semantic information from the Transformer using convolutional attention. The key difference between Multi-Scale Convolutional Attention (MSCA) and convolutional attention lies in our objective to replicate the attention mechanism using only convolutional operations, thereby completely eliminating the reliance on Transformers. This allows us to achieve a better trade-off between low cost and high performance. The distinction is also reflected in the structure. As shown in **Figure 3a**, MSCA incorporates multi-scale convolutional attention extraction, which includes both stripe convolution and conventional convolution. The former enhances efficiency and receptive field range, while the latter exploits the advantages of convolution in capturing local spatial semantic information. This design results in a high-performance attention module with low computation and memory cost. The specific formula is as follows:

$$x_{1,2,3} = \mathrm{Conv}\left(x, k_{v1,2,3}, \mathrm{padding} = (3,0)\right) \tag{3}$$

$$x_{1,2,3} = \mathrm{Attention}\left(x_{1,2,3}\right) \tag{4}$$

$$x_{1,2,3} = \mathrm{Conv}\left(x_{1,2,3}, k_{v1,2,3}^{T}, \mathrm{padding} = (3,0)\right) \tag{5}$$

$$x = \sum_{i=1}^{3} x_i \tag{6}$$

here, $x \in R^{n \times c \times h \times w}$ and $k_{v1,2,3}$ represent convolutional attention operations in the vertical, planar, and horizontal directions, respectively. The Attention includes reshaping the x dimensions and applying the Softmax normalization operation. Ultimately, x is transformed into an attention map with the same dimensions as the original.

The proposed Multi-Scale Convolutional Attention Mechanism employs multi-scale convolutional kernels to capture hierarchical spatial-contextual features across varying receptive fields, coupled with attention weighting to dynamically prioritize discriminative regions while suppressing redundant information, achieving adaptive

feature emphasis with minimal computational overhead This design enhances multi-scale feature extraction and adaptability for fine-grained tasks (e.g., medical imaging or IoT traffic classification) while maintaining computational efficiency through grouped convolutions, structural re-parameterization, and cross-scale interactions.

The multi-scale convolutional attention mechanism enhances feature extraction by integrating multi-scale convolutional kernels with attention weighting, capturing diverse spatial details while maintaining computational efficiency. In contrast, CFBlock focuses on contextual refinement using dilated convolutions, and Efficient Spatial Reduction Attention (ESRA) prioritizes computational efficiency through spatial reduction, making them less effective for preserving multi-scale feature richness.

3.3. Latent multi-conv-former generator

Our generator consists of a three-layer encoder-decoder structure, which includes a series of convolutional layers, normalization layers, and activation functions, focusing on capturing local spatial information of the image during the cross-modal generation process. For the encoder, the input consists of the source modality image x and the corresponding feature vector f. In the bottleneck layer, we employ nine MCF Blocks for long-range dependency modeling, allowing for better capture of the structural semantic information of medical images while maintaining low memory load. For the decoder, similar to [30,46], we use Adaptive Instance Normalization (AdaIN) to guide the generation of the target modality image using the feature vector. A detailed introduction to the generator loss function can be found in Section 3.4.

3.4. Discriminator and loss functions

We adopt the discriminator architecture from conditional PatchGAN [47] to provide the adversarial loss, thereby improving the performance of the generator and enhancing the quality of the generated images. The task of discriminator is to distinguish between real and generated images, while the generator aims to produce images that are as indistinguishable as possible to the discriminator. The specific formula is as follows:

$$\mathcal{L}_{\mathcal{D}} = \frac{1}{2}\lambda_{\text{adv}}[\mathcal{L}_{GAN}\left(D\left(x^{\text{real}}\right),\text{True}\right) + \mathcal{L}_{GAN}\left(D\left(x^{\text{fake}}\right),\text{False}\right)] \tag{7}$$

here, D represents the discriminator, x^{real} denotes the real image samples, and x^{fake} refers to the generated image samples. The parameter λ_{adv} serves as the weight coefficient for the loss, while $\mathcal{L}_{\text{GAN}}$ utilizes binary cross-entropy (BCE) loss.

The first term of the generator loss is the adversarial loss, which is expressed by the following formula:

$$\mathcal{L}_{adv} = \mathcal{L}_{GAN}\left(D\left(G\left(x^{real}, f\right)\right),\text{True}\right) \tag{8}$$

here, $G\left(x^{\text{real}}, f\right)$ represents the image generated by the generator, and f denotes the feature vector corresponding to the source modality image.

The second term represents the pixel-level loss, which uses the L1 loss to measure the difference between the synthesized image and the real samples. The specific

formula is as follows:

$$\mathcal{L}_{pix} = E_{x^{real},y,x^{fake}}\left[||G\left(x^{\text{real}}, f\right) - y||_1\right] \tag{9}$$

here, y represents the target modality image.

The third term is the cycle consistency loss, which further captures the mapping relationship between the two domains through reverse generation supervision. The specific formula is as follows:

$$\mathcal{L}_{cyc} = E_{x^{real},y,x^{fake}}\left[||G\left(x^{fake}, \hat{f}\right) - x^{real}||_1\right] \tag{10}$$

here, $\hat{f}$ represents the feature vector corresponding to the target modality image.

The fourth term is the feature consistency loss, which similarly ensures better domain consistency by reverse generating the feature vectors through the feature encoder. The specific formula is as follows:

$$\mathcal{L}_{fea} = E_{x^{real},y,x^{fake}}\left[||\hat{f} - FE\left(G\left(x^{real}, f\right)\right)||_1\right] \tag{11}$$

The fifth term is the orthogonal loss. To decouple the distinct styles of the two modality images as much as possible, an orthogonal form is used to make the vectors as independent as possible. The specific formula is as follows:

$$\mathcal{L}_{ort} = E_{x^{real},y,x^{fake}}[(f_{\text{wm}} \cdot f_{\text{w/o}})^2 + (f_{\text{tm}} \cdot f_{\text{w/o}})^2] \tag{12}$$

Therefore, the total loss function of the generator is expressed as follows:

$$\mathcal{L}_G = \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{pix} \cdot \mathcal{L}_{pix} + \lambda_{cyc} \cdot \mathcal{L}_{cyc} + \lambda_f \cdot \left(\mathcal{L}_{fea} + \mathcal{L}_{ort}\right) \tag{13}$$

here, λ_{adv}, λ_{pix}, λ_{cyc}, and λ_{f} are the weights associated with the corresponding losses.

4. Experiments and results

4.1. Datasets

We use the Spinal Disease dataset and the Multi-Modal Brain Tumor Segmentation Challenge 2020 (BraTS2020) dataset [48] to validate the effectiveness of the proposed method.

The Spinal Disease dataset contains imaging data from 300 patients with spinal disorders. However, due to the data being sourced from different hospitals and devices, there are variations in modality types and descriptions. To ensure consistency and enable experimental analysis, we divided the dataset into 200 training samples, 50 validation samples, and 50 test samples. Subsequently, based on the modality sequence descriptions in the Digital Imaging and Communications in Medicine (DICOM) files, we manually selected cases that included T1-weighted, T2-weighted, and Short-TI Inversion Recovery (STIR) imaging sequences, and processed them along the sagittal plane. As a result, the final number of valid 2D images for each modality was 615 for the training set, 165 for the validation set, and 146 for the test set.

The BraTS2020 dataset includes T1-weighted, T2-weighted, and T2 Fluid-Attenuated Inversion Recovery (FLAIR) imaging sequences. As a publicly available dataset, the skull has been removed, and all images are uniformly registered to the

same anatomical template. From this dataset, we selected data from 55 patients, with 25 used for the training set, 10 for the validation set, and 20 for the test set. For each patient, 100 two-dimensional slices were extracted from the axial plane.

All image slices in the experiment were uniformly resized to a standard 256 × 256 pixels, and the original intensity values were normalized to the range [0,1].

4.2. Evaluation metrics

To quantitatively assess the performance of the model, we utilized Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM). Among these, PSNR is a commonly used metric for evaluating image reconstruction quality, as it quantifies the signal-to-noise ratio between the original and reconstructed images. The formula is as follows:

$$\mathrm{PSNR} = 10 \cdot \log_{10}\left(\frac{R^2}{\mathrm{MSE}}\right) \tag{14}$$

here, R denotes the dynamic range of the image (e.g., for an 8-bit image, $R = 255$), and MSE represents the Mean Squared Error, which is calculated as follows:

$$\mathrm{MSE} = \frac{1}{m \cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n}\big(I(i,j) - K(i,j)\big)^2 \tag{15}$$

The Structural Similarity Index (SSIM) is used to evaluate the structural similarity between images, primarily by comparing luminance, contrast, and structural information to assess image quality. Its calculation is given by:

$$\mathrm{SSIM}(x,y) = \frac{\big(2\mu_x\mu_y + C_1\big)\big(2\sigma_{xy} + C_2\big)}{\big(\mu_x^2 + \mu_y^2 + C_1\big)\big(\sigma_x^2 + \sigma_y^2 + C_2\big)} \tag{16}$$

here, μ_x and μ_y represent the mean values of images x and y, respectively, while σ_{x}^2 and σ_{y}^2 denote their variances. σ_{xy} is the covariance between the two images, and C_1 and C_2 are constants used to stabilize the computation.

4.3. Comparison methods

We compare the proposed method with several classical and state-of-the-art approaches, including CycleGAN [20], pix2pix [48], L2M-GAN [30], PTNet [43], and ResViT [12]. All these methods are publicly available, and we trained them according to the provided settings until convergence.

4.4. Implementation details

We set the number of training epochs to 150, which matches the total training epochs of ResViT. The learning rate was set to 2e-4, with linear decay applied after 75 epochs. The hyperparameters of the Adam optimizer were set to $\beta_1 = 0.5$ and$\beta_2 = 0.999$. The hyperparameter values were $\lambda_{\mathrm{adv}} = 1$, $\lambda_{\mathrm{pix}} = 10$, $\lambda_{\mathrm{cyc}} = 1$, and $\lambda_{\mathrm{f}} = 100$. The experiments were run on a single NVIDIA GeForce RTX 4090.

4.5. Synthesis results

(1) Spinal Disease: We compare the generative performance metrics of our method with GAN-based and Transformer-based methods on six tasks: T2 → T1, T1 → T2, STIR → T1, T1 → STIR, STIR → T2, and T2 → STIR, as shown in **Table 1**. Except for the T2 → STIR and T2 → T1 tasks, CALF-GAN achieves the best performance on all other tasks. Given the parameter efficiency of CALF-GAN and its superior performance in SSIM, we conclude that the method demonstrates strong structural modeling capabilities, particularly on datasets with scarce and low-quality samples. This is attributed to the proposed multi-scale convolutional attention module, which effectively combines the advantages of contextual relationships and convolutional local precision.

Table 1. Performance for multi-contrast MRI translation tasks in spinal disease. psNR (dB) and SSIM are listed as mean ± std across the test set. boldface marks the top-performing model in each task. underline marks the second-performing model in each task.

	T2 → T1		**T1 → T2**		**STIR → T1**	
	PSNR ↑	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
CycleGAN	30.50 ± 0.634	0.648 ± 0.122	30.62 ± 0.661	0.618 ± 0.117	30.01 ± 0.741	0.521 ± 0.112
Pix2pix	30.41 ± 0.652	0.599 ± 0.109	30.50 ± 0.650	0.576 ± 0.127	30.18 ± 0.549	0.544 ± 0.079
L2M-GAN	30.42 ± 0.581	0.573 ± 0.084	30.33 ± 0.641	0.543 ± 0.102	30.02 ± 0.558	0.474 ± 0.085
PTNet	30.27 ± 0.557	0.609 ± 0.105	30.37 ± 0.474	0.598 ± 0.103	30.05 ± 0.491	0.541 ± 0.079
ResViT	30.48 ± 0.598	0.626 ± 0.103	30.65 ± 0.675	0.611 ± 0.126	30.18 ± 0.652	0.550 ± 0.084
CALF-GAN	30.49 ± 0.638	0.652 ± 0.105	30.68 ± 0.642	0.639 ± 0.118	30.28 ± 0.705	0.589 ± 0.090
	T1 → STIR		STIR → T2		T2 → STIR	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
CycleGAN	30.38 ± 0.671	0.511 ± 0.097	30.26 ± 0.741	0.544 ± 0.111	30.65 ± 0.751	0.551 ± 0.103
Pix2pix	30.20 ± 0.807	0.484 ± 0.091	30.35 ± 0.608	0.535 ± 0.104	30.41 ± 0.791	0.504 ± 0.092
L2M-GAN	29.65 ± 1.084	0.453 ± 0.102	30.16 ± 0.775	0.481 ± 0.109	30.44 ± 0.836	0.504 ± 0.097
PTNet	30.31 ± 0.650	0.519 ± 0.083	30.14 ± 0.529	0.540 ± 0.093	30.49 ± 0.672	0.568 ± 0.095
ResViT	30.59 ± 0.854	0.538 ± 0.094	30.38 ± 0.586	0.559 ± 0.099	30.82 ± 0.705	0.561 ± 0.096
CALF-GAN	30.80 ± 0.761	0.564 ± 0.088	30.45 ± 0.574	0.566 ± 0.107	30.71 ± 0.803	0.554 ± 0.101

Representative image results are shown in **Figure 4**. As indicated in **Figure 4a**, our method better captures the variations in signal intensities, such as those in the intervertebral discs, across different modalities. Although the intervertebral discs generated by Pix2pix are also relatively bright, the images generated by our method are overall closer to the target modality images and exhibit less noise compared to Pix2pix. As shown in **Figure 4b**, when zooming in on the third section of the intervertebral disc, CALF-GAN more accurately captures the disc bulging to the left. Furthermore, our method avoids the issue observed in CycleGAN and PTNet, where the morphology of the third section of the intervertebral disc is closer to the source modality image rather than the target modality image. Overall, compared to baseline methods, CALF-GAN demonstrates more reliable performance in describing tissue structure details and generating overall modality style.

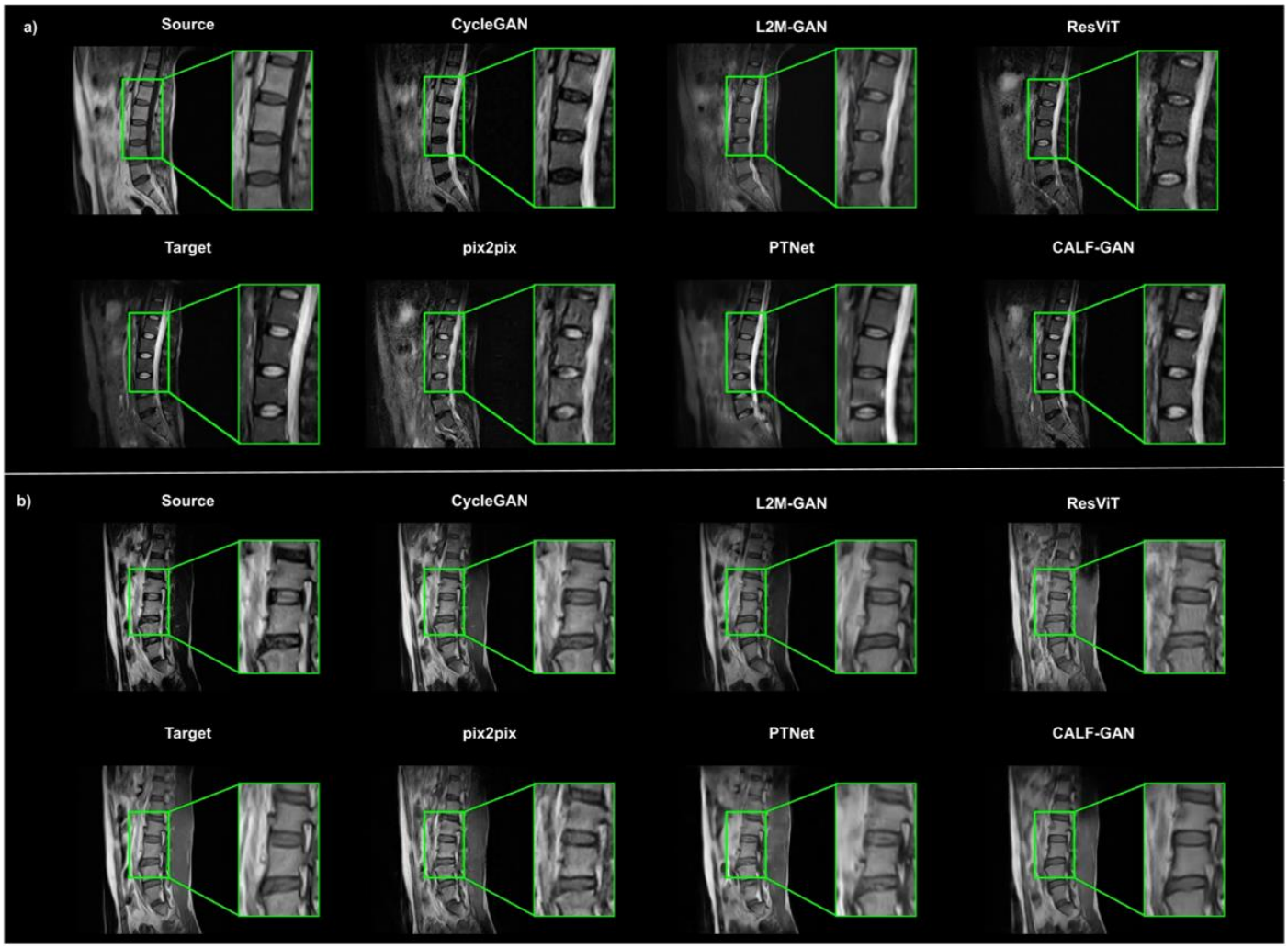

Figure 4. Qualitative results of CALF-GAN and the comparison methods on the spinal disease dataset **(a)** for the T1STIR task; **(b)** for the T2T1 task.

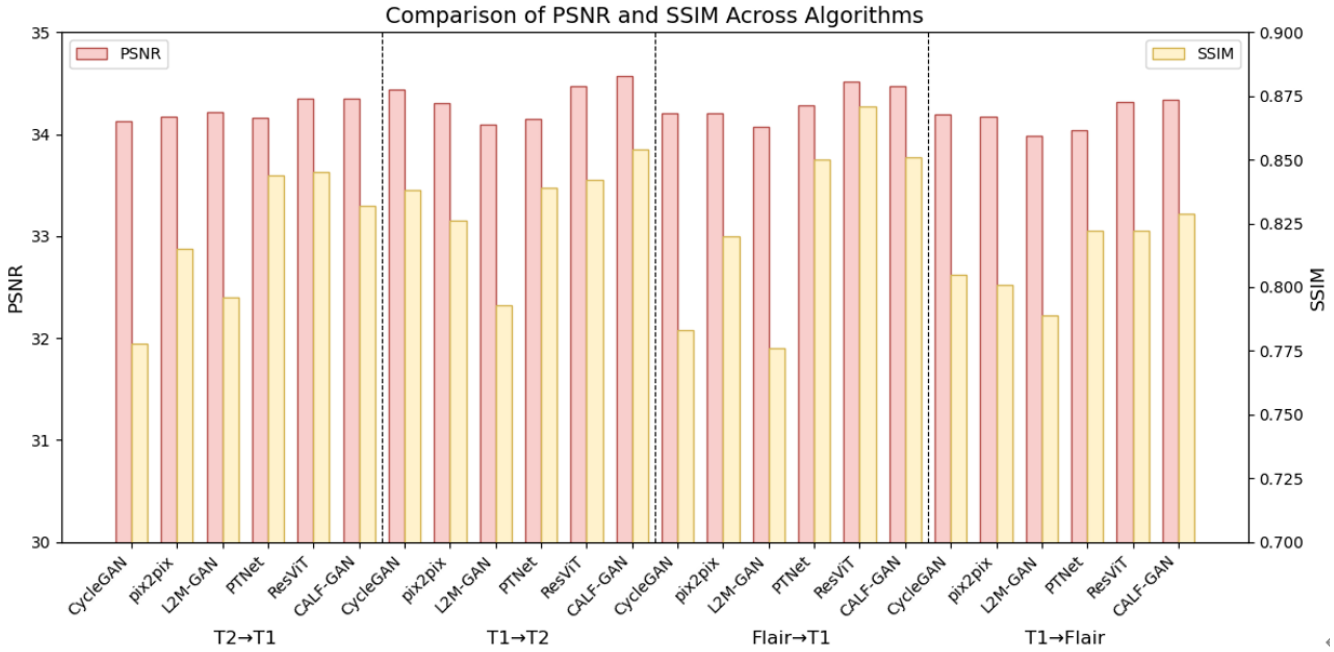

Figure 5. Quantitative results of CALF-GAN and comparison methods on the BraTS2020 dataset.

BraTS2020: We compare the generative performance metrics of CALF-GAN with GAN-based and Transformer-based methods on four tasks: T2 → T1, T1 → T2, FLAIR → T1, and T1 → FLAIR, as shown in **Figure 5**. CALF-GAN maintains an advantage in the T1 → T2 and T1→ FLAIR tasks, while in the remaining tasks, it outperforms other GAN methods and performs similarly to Transformer-based methods. We attribute this to the inherent advantage of Transformer methods on datasets with a sufficiently large number of samples. Overall, CALF-GAN provides a more favorable trade-off between computational cost and generative performance.

We select representative images from tasks where CALF-GAN outperforms the competing methods (T1 → FLAIR) and performs similarly to the competing methods (T2 → T1), as shown in **Figure 6**. First, in terms of local spatial structure semantics, CALF-GAN demonstrates superior performance, particularly in details of the glioma region and brain structural textures. Specifically, for the T1 → FLAIR task, the pixel intensity values generated by L2M-GAN lie between those of the source and target modalities. Images generated by Pix2pix and PTNet are relatively blurry and inaccurate. CycleGAN and ResVit exhibit advantages in the localization of diffuse lesion borders and complex-texture lesions, while CALF-GAN performs well in both areas. For the T2→T1 task, except for CycleGAN, the images generated by other methods are relatively accurate. Compared to competing methods, CALF-GAN also demonstrates finer local spatial texture generation ability, although some noise artifacts are present. Overall, the experimental results demonstrate that CALF-GAN exhibits competitive generative performance even when trained on datasets with a sufficient number of samples.

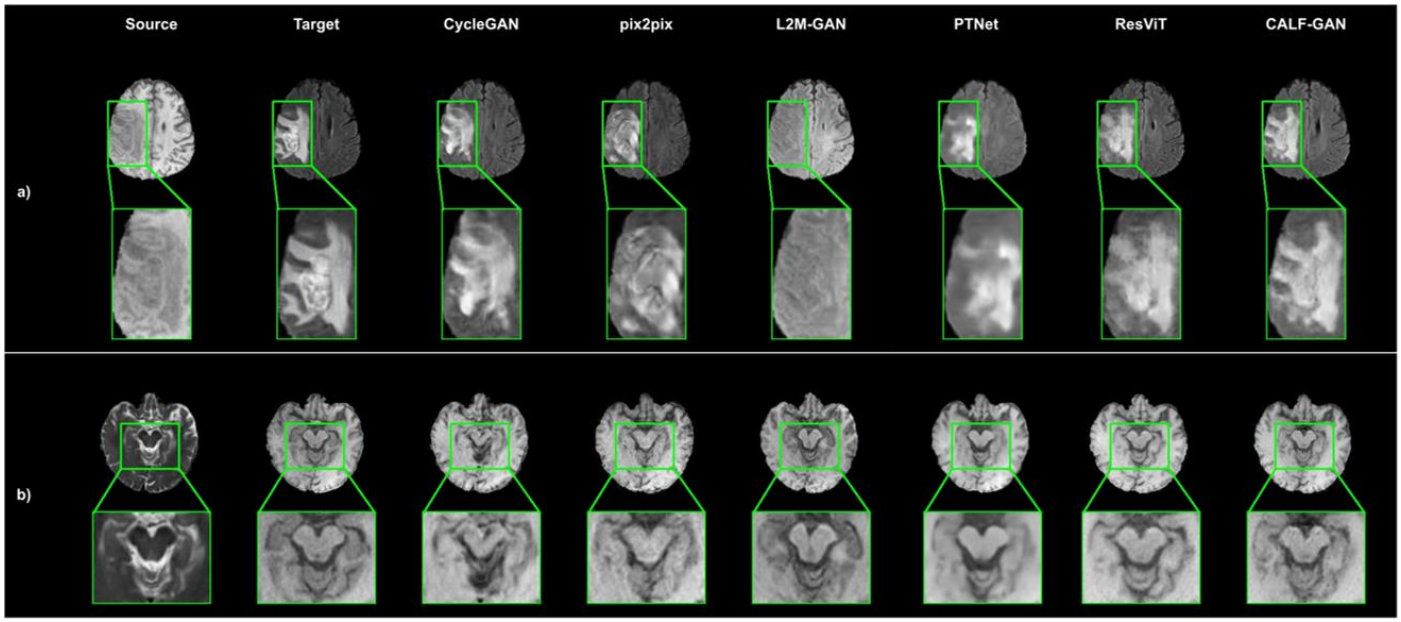

Figure 6. Qualitative results of CALF-GAN and comparison methods on the BraTS2020 dataset **(a)** T1 → FLAIR task; **(b)** T2 → T1 task.

4.6. Model complexity

Model complexity is an important consideration in environments with limited computational resources. **Table 2** presents a comparison of CALF-GAN and the

competing methods in terms of generator parameter count, total model parameters, and GPU memory usage. In terms of generator parameter count, CALF-GAN significantly outperforms the other methods. Regarding total model parameters, CALF-GAN is comparable to CycleGAN and PTNet, yet still superior to the other methods. In terms of GPU memory usage, CALF-GAN requires significantly less memory than the other competing methods, except for Pix2pix. Considering the superior generative performance of CALF-GAN, we conclude that our approach achieves a more favorable trade-off between generative quality and computational efficiency.

Table 2. Comparison of Parameters and Memory Load with an input size of (256, 256).

	CycleGAN	**Pix2pix**	**L2M-GAN**	**PTNet**	**ResViT**	**CALF-GAN**
Generator Model Complexity (M)	22.74	54.41	33.89	27.69	123.44	14.79
Total Model Complexity (M)	28.26	57.78	87.40	30.80	126.20	26.31
Total GPU VRAM Usage (GB)	9.50	2.83	14.68	5.34	3.66	2.81

4.7. Ablation studies

We conducted ablation experiments to demonstrate the effectiveness of each module in CALF-GAN, as shown in **Table 3**. First, the baseline is based on the proposed Latent Multi-Conv-Former (LMCF) Generator, where all MCF Blocks are replaced with standard convolutional blocks. Next, we investigate the role of latent feature guidance in the model, where SE refers to the Style Encoder block proposed in [30], representing the introduced latent space attribute decomposition mechanism, and FE refers to the improved feature encoding block we propose. We then explore the added value of the MCF Block to the model's performance. Finally, the improvement in SSIM for our method is relatively significant, which is consistent with the performance observed in the experiments. In summary, each of the proposed modules contributes to the enhancement of the model's generative performance.

Table 3. Quantitative evaluation results by comparing our full model with its ablated versions on the spinal disease dataset.

Models	**T1 → T2**		**T2 → T1**	
	PSNR ↑	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
Baseline	30.55 ± 0.604	0.598 ± 0.100	30.41 ± 0.507	0.611 ± 0.086
Baseline + SE	30.57 ± 0.597	0.600 ± 0.104	30.41 ± 0.630	0.613 ± 0.101
Baseline + FE	30.64 ± 0.710	0.606 ± 0.108	30.45 ± 0.559	0.623 ± 0.098
Baseline + MCF Block	30.67 ± 0.640	0.625 ± 0.107	30.48 ± 0.574	0.639 ± 0.103
Baseline + FE + MCF Block	30.68 ± 0.642	0.639 ± 0.118	30.49 ± 0.638	0.652 ± 0.105

To further investigate the actual impact of the proposed multi-scale convolutional attention mechanism in the model, we utilize Grad-CAM [49] to visualize the attention heatmaps, illustrating the enhancement of the model's focus on specific regions. We apply Grad-CAM to generate heatmaps for the same convolutional layer (the final convolution block in the bottleneck layer) in both the baseline method and CALF-GAN. As shown in **Figure 7**, compared to the baseline, the multi-scale convolutional

attention in the MCF Block enables our model to better perceive the regions of interest (ROI) in the spine.

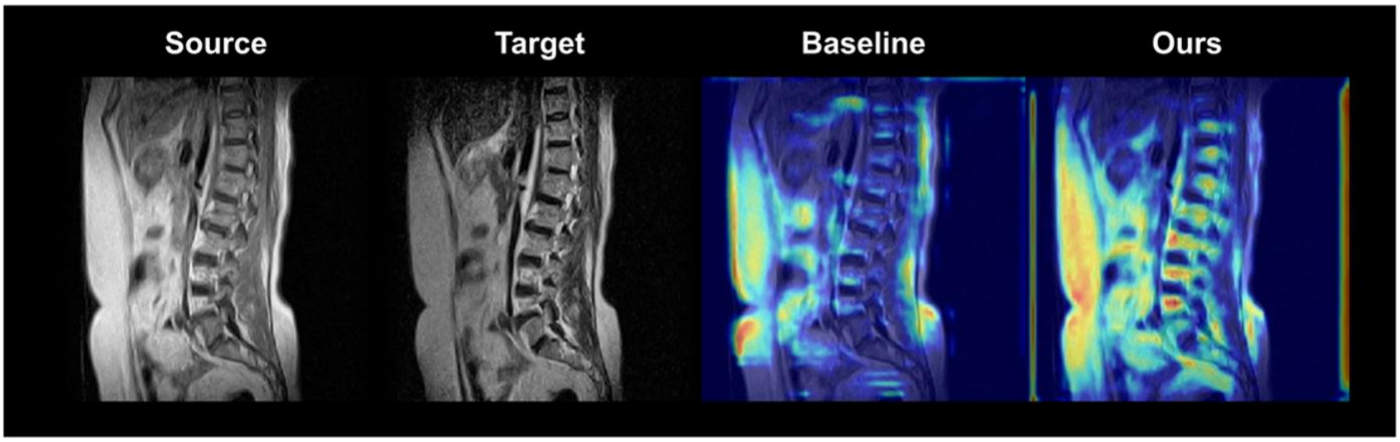

Figure 7. Attention heatmaps demonstrating the effect of the MCF block.

5. Discussion

The absence of specific modality images and the scarcity of high-quality medical image samples have become common challenges in clinical diagnosis, biomechanics research, and artificial intelligence model training. Although technologies such as Transformers have made significant progress in cross-modal generation of high-quality medical images, these methods still exhibit notable limitations in research environments with scarce samples, low-quality data, and limited computational resources. To address these challenges, this paper proposes a generative approach that achieves a good balance between low computational cost and high generative performance. Extensive experiments have demonstrated that the images generated by this method possess sufficient realism to meet the practical application requirements of biomechanics research.

Through extensive quantitative experiments, we demonstrate that the proposed method achieves competitive results despite having significantly fewer parameters than the comparison methods. To ensure the fairness of comparison with L2M-GAN, we transfer our method to a unified generator head model, named CALF-GAN-uni, which is designed to perform six spine cross-modal synthesis tasks simultaneously. Specifically, the unified generator head refers to using a single generator to synthesize all six modalities. This approach improves training efficiency and reduces computational costs, but it may lead to certain tasks converging earlier, making it challenging to achieve optimal performance across all tasks simultaneously. The experimental results (as shown in **Table 4**) indicate that, even when transferred to a unified generator model, our method still demonstrates excellent performance. It not only surpasses L2M-GAN but also achieves generative performance comparable to Transformer-based methods, thus proving the versatility of our approach. However, there is still room for improvement in the selection of loss functions and the discriminator architecture. Specifically, the choice of loss functions (such as L1 and L2 losses) and further optimization of the discriminator architecture or loss functions are important directions for future work to enhance CALF-GAN's performance.

Table 4. Performance for CALF-GAN-uni in spinal disease. PSNR (dB) and SSIM are listed as mean ± std across the test set.

	T2 → T1		**T1 → T2**		**STIR → T1**	
	PSNR ↑	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**	**PSNR ↑**	**SSIM ↑**
L2M-GAN	30.42 ± 0.581	0.573 ± 0.084	30.33 ± 0.641	0.543 ± 0.102	30.02 ± 0.558	0.474±0.085
ResViT	30.48 ± 0.598	0.626 ± 0.103	30.65 ± 0.675	0.611 ± 0.126	30.18 ± 0.652	0.550 ± 0.084
CALF-GAN-uni	30.56 ± 0.773	0.650 ± 0.118	30.55 ± 0.683	0.623 ± 0.127	30.06 ± 0.722	0.534 ± 0.092
	T1 → STIR		STIR → T2		T2 → STIR	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
L2M-GAN	29.65 ± 1.084	0.453 ± 0.102	30.16 ± 0.775	0.481 ± 0.109	30.44 ± 0.836	0.504 ± 0.097
ResViT	30.59 ± 0.854	0.538 ± 0.094	30.38 ± 0.586	0.559 ± 0.099	30.82 ± 0.705	0.561 ± 0.096
CALF-GAN-uni	30.53 ± 0.783	0.535 ± 0.085	30.33 ± 0.771	0.553 ± 0.119	30.80 ± 0.757	0.571 ± 0.103

Additionally, the spinal dataset is derived from a publicly available collection that integrates data from multiple hospitals. After manually selecting paired spinal images with three modalities, we retained some low-quality, noisy images to evaluate the model's generative capability on lower-quality samples. This decision was also made considering the scarcity of spinal samples. Furthermore, to demonstrate the generative quality on standard medical image samples, we used the publicly available BraTS brain dataset. The experimental results show that our method exhibits superior generative performance on low-quality sample datasets while also achieving competitive results on normal sample data. Specifically, as shown in **Figures 4** and **6**, CALF-GAN demonstrates strong capabilities in global semantic modeling and detailed texture generation, which can be attributed to the effective contribution of the proposed modality feature guidance module and multi-scale convolutional attention mechanism.

To further visualize the enhancement provided by the proposed method to the model, we use Grad-CAM to display the effect of the multi-scale convolutional attention mechanism in the form of attention heatmaps, as shown in **Figure 7**. Compared to the baseline method, our approach not only emphasizes the highlighted regions of the fat areas but also effectively captures the structural semantic information of the spine, thereby improving the overall quality of the generated images. This improvement is also reflected in the superior SSIM performance, as shown in **Table 1**.

However, CALF-GAN still has certain limitations. First, the introduction of the latent feature guidance module reduces computational efficiency compared to the baseline method. Although we have designed it with lower memory costs, the incorporation of the FE and FC modules still introduces some additional computational overhead. Second, there remains considerable room for improvement in the generation quality of non-critical regions, such as soft tissues. Specifically, as shown in **Figure 4b**, the soft tissues adjacent to the spine exhibit an over-smoothing effect. This may be due to the attention mechanism focusing the regions of interest (ROI) on the spinal structure and highlighted fat-water signal modalities, which leads to a decline in the generation performance for non-critical areas. Therefore, future work should focus on further enhancing the generative realism of these regions to

improve overall image quality.

Furthermore, biomechanics research based on medical imaging will provide stronger support for personalized medicine. Enhancing the clinical applicability of generated images and gaining the recognition of clinicians are crucial directions for future research. By analyzing patients' medical images and conducting biomechanical modeling, it is possible to predict individual responses to specific treatment methods, enabling the development of the most suitable treatment plans for each patient. When generated images are used for high-precision biomechanical research and clinical decision support, they must meet more stringent standards, such as maintaining the continuity and consistency of pixel intensity in spinal images. Therefore, advancing research in this area not only helps to expand the application scope of generated images but also increases their clinical value and practical significance.

6. Conclusion

In this paper, we propose a latent feature-guided generative adversarial model based on multi-scale convolutional attention (CALF-GAN). The model consists of a latent space attribute separation module and a generative module based on multi-scale convolutional attention. The attribute separation module is designed to extract global semantic information from image modalities, while the generative module, which simulates the Transformer architecture using only convolution operations, achieves an effective balance between local convolutional precision, long-range dependency modeling, and low computational cost. Extensive experimental results validate the effectiveness of CALF-GAN in overcoming the aforementioned limitations. In addition, the high-resolution and high-fidelity medical images generated by this model exhibit significant potential for development and application in the field of biomechanics research. This study provides precise geometric modeling and dynamic analysis capabilities for biomechanics research, aiding in uncovering the complexities of human mechanical behavior and advancing the development of precision medicine.

In the future, advancements in artificial intelligence and big data technologies will further integrate medical image recognition with biomechanics, expanding the boundaries of medicine, engineering, and biological sciences to enable more cross-disciplinary innovative applications. We will continue to explore alternative loss functions and refinement techniques to enhance the realism of non-critical regions.

Author contributions: Conceptualization, XZ and YW; methodology, XZ; software, XZ; validation, XZ and YW; formal analysis, XZ; investigation, XZ; resources, XZ; data curation, XZ; writing—original draft preparation, XZ; writing—review and editing, XZ; visualization, XZ; supervision, XZ; project administration, XZ; funding acquisition, YW. All authors have read and agreed to the published version of the manuscript.

Ethics approval: Not applicable.

Conflict of interest: The authors declare no conflict of interest.

References

1. Wang L, Zhu Z. Applications and challenges of artificial intelligence-driven 3D vision in biomedical engineering: A biomechanics perspective. Molecular & Cellular Biomechanics. 2025; 22(2): 1006. doi: 10.62617/mcb1006
2. Barkaoui A, Ait Oumghar I, Ben Kahla R. Review on the use of medical imaging in orthopedic biomechanics: finite element studies. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 2021; 9(5): 535–554. doi: 10.1080/21681163.2021.1888317
3. Seyedpour SM, Nabati M, Lambers L, et al. Application of Magnetic Resonance Imaging in Liver Biomechanics: A Systematic Review. Frontiers in Physiology. 2021; 12. doi: 10.3389/fphys.2021.733393
4. Galbusera F, Cina A, Panico M, et al. Image-based biomechanical models of the musculoskeletal system. European Radiology Experimental. 2020; 4(1). doi: 10.1186/s41747-020-00172-3
5. Zhang J, Sun K, Yang J, et al. A generalized dual-domain generative framework with hierarchical consistency for medical image reconstruction and synthesis. Communications Engineering. 2023; 2(1). doi: 10.1038/s44172-023-00121-z
6. Yang H, Zhou T, Zhou Y, et al. Flexible Fusion Network for Multi-Modal Brain Tumor Segmentation. IEEE Journal of Biomedical and Health Informatics. 2023; 27(7): 3349–3359. doi: 10.1109/jbhi.2023.3271808
7. Mahapatra D, Bozorgtabar B, Ge Z, et al. GANDALF: Graph-based transformer and Data Augmentation Active Learning Framework with interpretable features for multi-label chest Xray classification. Medical Image Analysis. 2024; 93: 103075. doi: 10.1016/j.media.2023.103075
8. van Herten RLM, Hampe N, Takx RAP, et al. Automatic Coronary Artery Plaque Quantification and CAD-RADS Prediction Using Mesh Priors. IEEE Transactions on Medical Imaging. 2024; 43(4): 1272–1283. doi: 10.1109/tmi.2023.3326243
9. Cao H, Wang Y, Chen J, et al. Swin-unet: Unet-like pure transformer for medical image segmentation. In: Proceedings of the European conference on computer vision; 2022. pp. 205–218.
10. Thukral BB. Problems and preferences in pediatric imaging. Indian Journal of Radiology and Imaging. 2015; 25(04): 359–364. doi: 10.4103/0971-3026.169466
11. Krupa K, Bekiesińska-Figatowska M. Artifacts in Magnetic Resonance Imaging. Polish Journal of Radiology. 2015; 80: 93–106. doi: 10.12659/pjr.892628
12. Dalmaz O, Yurt M, Cukur T. ResViT: Residual Vision Transformers for Multimodal Medical Image Synthesis. IEEE Transactions on Medical Imaging. 2022; 41(10): 2598–2614. doi: 10.1109/tmi.2022.3167808
13. Liu J, Pasumarthi S, Duffy B, et al. One Model to Synthesize Them All: Multi-Contrast Multi-Scale Transformer for Missing Data Imputation. IEEE Transactions on Medical Imaging. 2023; 42(9): 2577–2591. doi: 10.1109/tmi.2023.3261707
14. Luo Y, Nie D, Zhan B, et al. Edge-preserving MRI image synthesis via adversarial network with iterative multi-scale fusion. Neurocomputing. 2021; 452: 63–77. doi: 10.1016/j.neucom.2021.04.060
15. Chartsias A, Joyce T, Giuffrida MV, et al. Multimodal MR Synthesis via Modality-Invariant Latent Representation. IEEE Transactions on Medical Imaging. 2018; 37(3): 803–814. doi: 10.1109/tmi.2017.2764326
16. Sevetlidis V, Giuffrida MV, Tsaftaris SA. Whole image synthesis using a deep encoder-decoder network. In: Simulation and Synthesis in Medical Imaging: First International Workshop, SASHIMI 2016. Springer; 2016. pp. 127–137.
17. Huang P, Li D, Jiao Z, et al. Common feature learning for brain tumor MRI synthesis by context-aware generative adversarial network. Medical Image Analysis. 2022; 79: 102472. doi: 10.1016/j.media.2022.102472
18. Ang SP, Lam Phung S, Field M, et al. An Improved Deep Learning Framework for MR-to-CT Image Synthesis with a New Hybrid Objective Function. 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). 2022: 1–5. doi: 10.1109/isbi52829.2022.9761546
19. Cao B, Zhang H, Wang N, et al. Auto-GAN: Self-Supervised Collaborative Learning for Medical Image Synthesis. Proceedings of the AAAI Conference on Artificial Intelligence. 2020; 34(07): 10486–10493. doi: 10.1609/aaai.v34i07.6619
20. Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision; 2017. pp. 2223–2232.
21. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Advances in neural information processing systems. 2017; 30.
22. Li Y, Zhou T, He K, et al. Multi-Scale Transformer Network with Edge-Aware Pre-Training for Cross-Modality MR Image Synthesis. IEEE Transactions on Medical Imaging. 2023; 42(11): 3395–3407. doi: 10.1109/tmi.2023.3288001
23. Chen J, Lu Y, Yu Q, Luo X, et al. Transunet: Transformers make strong encoders for medical image segmentation. arXiv;

2021.
24. Zhao B, Cheng T, Zhang X, et al. CT synthesis from MR in the pelvic area using Residual Transformer Conditional GAN. Computerized Medical Imaging and Graphics. 2023; 103: 102150. doi: 10.1016/j.compmedimag.2022.102150
25. Sun H, Wen Y, Feng H, et al. Unsupervised Bidirectional Contrastive Reconstruction and Adaptive Fine-Grained Channel Attention Networks for image dehazing. Neural Networks. 2024; 176: 106314. doi: 10.1016/j.neunet.2024.106314
26. Xu Z, Wu D, Yu C, et al. SCTNet: Single-Branch CNN with Transformer Semantic Information for Real-Time Segmentation. Proceedings of the AAAI Conference on Artificial Intelligence. 2024; 38(6): 6378–6386. doi: 10.1609/aaai.v38i6.28457
27. Li X, Qin X, Huang C, et al. SUnet: A multi-organ segmentation network based on multiple attention. Computers in Biology and Medicine. 2023; 167: 107596. doi: 10.1016/j.compbiomed.2023.107596
28. Shen Y, Gu J, Tang X, et al. Interpreting the Latent Space of GANs for Semantic Face Editing. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2020. pp. 9243–9252.
29. Zhu J, Shen Y, Zhao D, Zhou B. In-domain gan inversion for real image editing. In: Proceedings of the European conference on computer vision; 2020. pp. 592–608.
30. Yang G, Fei N, Ding M, et al. L2M-GAN: Learning to Manipulate Latent Space Semantics for Facial Attribute Editing. In: Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021. pp. 2951–2960.
31. Masutani EM. Deep Learning Image Synthesis for MRI: From Super-Resolution to Cardiovascular Biomechanics. University of California, San Diego; 2022.
32. Peng X, Tang L. Biomechanics analysis of real-time tennis batting images using Internet of Things and deep learning. The Journal of Supercomputing. 2021; 78(4): 5883–5902. doi: 10.1007/s11227-021-04111-w
33. Shi Y, Ma S, Zhao Y, et al. A Physics-Informed Low-Shot Adversarial Learning for sEMG-Based Estimation of Muscle Force and Joint Kinematics. IEEE Journal of Biomedical and Health Informatics. 2024; 28(3): 1309–1320. doi: 10.1109/jbhi.2023.3347672
34. Sohail M, Riaz MN, Wu J, Long C, LiS. Unpaired multi- contrast mr image synthesis using generative adversarial networks. In: Proceedings of the International Workshop on Simulation and Synthesis in Medical Imaging; 2019. pp. 22–31.
35. Ge Y, Wei D, Xue Z, et al. Unpaired Mr to CT Synthesis with Explicit Structural Constrained Adversarial Learning. In: Proceedings of the 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019); 2019. pp. 1096–1099.
36. Dong X, Wang T, Lei Y, et al. Synthetic CT generation from non-attenuation corrected PET images for whole-body PET imaging. Physics in Medicine & Biology. 2019; 64(21): 215016. doi: 10.1088/1361-6560/ab4eb7
37. Yurt M, Dar SUH, Özbey M, et al. Semi-supervised learning of mutually accelerated mri synthesis without fully-sampled ground truths. arXiv; 2020.
38. Liang X, Chen L, Nguyen D, et al. Generating synthesized computed tomography (CT) from cone-beam computed tomography (CBCT) using CycleGAN for adaptive radiation therapy. Physics in Medicine & Biology. 2019; 64(12): 125002. doi: 10.1088/1361-6560/ab22f9
39. Emami H, Dong M, Nejad-Davarani SP, Glide- Hurst CK. Sa-gan: Structure-aware gan for organ-preserving synthetic ct generation. In: Proceedings of the Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference; 2021. pp. 471–481.
40. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv; 2020.
41. Perera S, Adhikari S, Yilmaz A. Pocformer: A Lightweight Transformer Architecture For Detection Of Covid-19 Using Point Of Care Ultrasound. In: Proceedings of the 2021 IEEE International Conference on Image Processing (ICIP). 2021. pp. 195–199.
42. Chen J, He Y, Frey EC, et al. Vit-v-net: Vision transformer for unsupervised volumetric medical image registra- tion. arXiv; 2021.
43. Zhang X, He X, Guo J, et al. Ptnet: A high-resolution infant mri synthesizer based on transformer. arXiv; 2021.
44. Fetty L, Bylund M, Kuess P, et al. Latent space manipulation for high-resolution medical image synthesis via the StyleGAN. Zeitschrift für Medizinische Physik. 2020; 30(4): 305–314. doi: 10.1016/j.zemedi.2020.05.001
45. Ioffe S. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv; 2015.
46. Choi Y, Uh Y, Yoo J, et al. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In: Proceedings of the 2020

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Published online June 2020. pp. 8188– 8197.
47. Isola P, Zhu JY, Zhou T, et al. Image-to-Image Translation with Conditional Adversarial Networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017: pp. 1125–1134.
48. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging. 2015; 34(10): 1993–2024. doi: 10.1109/tmi.2014.2377694
49. Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV); 2017. pp. 618–626. doi: 10.1109/iccv.2017.74