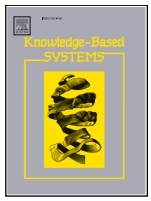
Knowledge-Based Systems 334 (2026) 115017

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



AttCL-GAN: Attentional contrastive learning-based generative adversarial network for modality completion of medical images

Zhenghua Xu [a,b,*], Jiaqi Tang [a,b], Dan Yao [a,b], Zhenzhen Wang [a,b], Thomas Lukasiewicz [c,d]

[a] State Key Laboratory of Intelligent Power Distribution Equipment and System, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China
[b] Hebei Engineering Research Center of Brain-Computer Intelligent Fusion Technology, School of Health Sciences and Biomedical Engineering, Hebei University of Technology, Tianjin, China
[c] Institute of Logic and Computation, Vienna University of Technology, Vienna, Austria
[d] Department of Computer Science, University of Oxford, Oxford, United Kingdom

ARTICLE INFO

Keywords:
Modality completion of medical images
Multi-modal image segmentation
Generative adversarial networks
Attention mechanism
Contrastive learning

ABSTRACT

Most multi-modal medical image analysis models require datasets to be modal-balanced. However, existing multi-modal medical datasets often suffer from modality missing problems due to practical factors. Therefore, in this study, we propose a novel generative model, AttCL-GAN, to remedy three issues encountered when applying the multi-domain model StarGAN v2 to medical images generation. Further by utilizing the proposed modality completion strategy, we achieve the data imputation under arbitrary modality missing scenarios. The advancements of AttCL-GAN are mainly threefold: Firstly, we integrate the idea of contrastive learning and attention mechanism to guide the network in learning domain-independent content during image translations using attention information from spatial and channel dimensions. Then, the attentional AdaIN generator is introduced into the network. By adding the refined features and original features together and feeding them into AdaIN, the generator's ability to generate tissue details is enhanced. Finally, to alleviate the problem of high inter-modal similarity of generated images, we propose a style code diversity loss, which increases the diversity of images by enlarging the Euclidean distance between codes of different modalities in the latent space. Extensive experimental results on a real-world multi-modal brain MRI dataset show that (i) The proposed AttCL-GAN significantly outperforms the state-of-the-art GAN-based data augmentation methods in both the generation and segmentation tasks in terms of all metrics; (ii) The proposed three advancements are all effective and essential for AttCL-GAN to achieve the superior performances in both tasks.

1. Introduction

By integrating information from different imaging modalities such as MRI, CT, PET, etc., multi-modal segmentation methods significantly improve model performances, rendering them adaptable to varied clinical scenarios [1–6]. For example, Chen et al. [5] proposed RFDCR, a two-stage framework that segments brain tumors and stroke lesions by jointly optimizing local and global features from multi-modal MRI. However, most multi-modal segmentation models require balanced modality data in the training set, meaning that each patient must have images from all considered modalities. This condition is difficult to achieve in clinical practice due to factors such as device independence and privacy concerns, with most multi-modal datasets facing situations where images from certain modalities are missing. Therefore, to address the modality missing issue in current multi-modal medical image datasets, utilizing cross-domain generative adversarial networks (GANs) for data augmentation has become a research hotspot.

By learning the mapping relationships between different modalities, achieving image translations from one modality to another, GANs have achieved numerous successful applications in cross-modal generation of medical images. This approach provides an effective data augmentation method for downstream medical analysis tasks. For example, to reduce the radiation exposure of CT and improve the limitations of traditional virtual imaging techniques, Gu et al. [7] proposed a MRI-to-CT generative method with structural perceptual supervision. High costs of equipment and various contraindications present significant challenges for those seeking PET scans. To address this issue, Li et al. [8] proposed the C2P-GAN, which is based on a fully convolutional transformer and

* Corresponding author.
E-mail address: zhenghua.xu@hebut.edu.cn (Z. Xu).

https://doi.org/10.1016/j.knosys.2025.115017
Received 18 August 2024; Received in revised form 16 October 2025; Accepted 28 November 2025
Available online 10 December 2025

residual blocks to perform cross-modal synthesis of PET images from CT images. Qin et al. [9] introduced style transfer into CGAN and proposed a CGAN-based model with hierarchical feature mapping and fusion, named ST-cGAN, to achieve cross-modal synthesis within MRI images. Although these methods successfully generate high-quality images from one modality to another, they are limited to A-to-B translations, such as MRI-to-PET. When faced with unseen modalities, these models require retraining to learn the new mappings. Therefore, to address the above issue, multi-domain generative models have been proposed.

StarGAN and StarGAN v2 are multi-domain generation models proposed by Choi et al [10,11], which can achieve arbitrary translations between multiple domains using a single network architecture. To preserve crucial structures while performing style transfer, Tom et al. [12] proposed a method that translated unstained images to realistically looking stained images by improving StarGAN. Peng et al. [13] applied the improved model of StarGAN v2 to augment a brain tumor dataset, proving that generated multi-modal images can enhance the performance of downstream segmentation tasks. In general, existing researches have demonstrated the successful application of StarGAN series in arbitrary image translations. However, it has also exposed the following issues when applied to medical imaging: (i) **Poor lesion generation capability**: Lesion samples are typically scarce and overshadowed by normal samples in medical imaging datasets. When using a normal reference, this imbalance may lead generators to overlook lesion generation. (ii) **Insufficient tissue detail synthesis**: Human tissues exhibit complex structures and diversity, such as cells and blood vessels. Multi-domain models may prioritize the transfer of stylistic patterns while neglecting the accurate capture and restoration of intricate tissue structures. (iii) **High inter-modal similarity of results**: Although different modalities depict complementary aspects of the same anatomy, their inherent structural similarities may cause generative networks to struggle in producing sufficiently distinct outputs across modalities.

Consequently, given the context and issues mentioned above, we propose an Attentional Contrastive Learning-based Generative Adversarial Network (AttCL-GAN) for modality completion. The network adopts StarGAN v2 as the basic framework and addresses the issues of poor lesion generation capability, insufficient tissue detail synthesis and high inter-modal similarity of results respectively using the Attention-guided Contrastive Learning (ACL) module, Attentional AdaIN-based (AttAdaIN) generator and Style Code Diversity (SCD) loss, achieving arbitrary image translations and effective image generations among multi-modal medical images. Furthermore, based on the modality completion strategy we proposed, the model can perform data imputation on multi-modal datasets with various modality missing situations. This achieves the goal of increasing the utilization of imbalanced multi-modal images and improves the accuracy of downstream segmentation models.

Overall, compared to the vanilla StarGAN v2, AttnCL-GAN has made advancements in the following three aspects. Firstly, building upon the original StarGAN v2, we introduce the attention-guided contrastive learning module, which is implemented by a patch sampler and an attention network. The patch sampler is utilized to select anchor points,also called queries, their corresponding positives, and many negatives for computing contrastive loss. The attention network is employed to acquire an attention map guiding the selection of queries and the AdaIN layers in the generator. Through this loss, the generative network can better learn common information in different modalities while ignoring modality-dependent information, improving the generation of lesions.

Furthermore, the second advancement of AttCL-GAN lies in the attentional AdaIN-based generator. Specifically, the conventional AdaIN layer takes only content and style features as input, achieving style transfer in the feature space by passing mean and variance. In AttCL-GAN, however, the attention map obtained from the attention network, along with the original features of that layer, is added together and fed into the AdaIN layer of the decoder in generator. Following this operation, the information entering the AdaIN layer transforms from being single-dimensional to multi-dimensional, which empower the generator to enhance its synthesis of tissue details.

Finally, to alleviate the issue of high inter-modal similarity of results, we introduce a style code diversity loss on the generator. In the case of having N domains, the obtained $1 \times N$ style code contains the probabilities of predicting the input image as every domain. When the target domain is known, the loss is acquired by computing the Euclidean distances between the predicted probabilities for that domain and other non-target domains. The average of these distances is then defined as the SCD loss. Subsequently, by maximizing this loss, the model enhances the modal distinctiveness of results.

The contributions of this paper can be summarized as follows:

- We identified three issues arising when StarGAN v2 was applied to medical image generation and proposed a model named AttCL-GAN to alleviate these issues and achieve effective image translations among arbitrary multi-modal medical images.
- In AttCL-GAN, an attention-guided contrastive learning (ACL) module is first proposed to address the issue of poor lesion generation capability, consists of two new network components. Subsequently, to alleviate the problem of insufficient tissue detail synthesis, a attentional AdaIN-based (AttAdaIN) generator is further introduced, which uses content and spatial attention information to guide feature normalization. Finally, we add a style code diversity (SCD) loss to remedy the high inter-modal similarity in generated images, without adding extra model capacity.
- Furthermore, we also proposed a modality completion strategy, with which AttCL-GAN is employed for data imputation on multi-modal datasets with various modality missing scenarios.
- Extensive experimental studies are conducted on a real-world multi-modal brain MRI dataset, and the results show the following: (i) The proposed AttCL-GAN significantly outperforms the state-of-the-art GAN-based data augmentation methods in both the generation and segmentation tasks in terms of all metrics. (ii) The proposed three advancements are all effective and essential for AttCL-GAN to achieve the superior performances in both tasks.

2. Related work

Multi-domain image-to-image translation. The aim of image-to-image (I2I) translation is to learn a mapping function $G : X \rightarrow Y$ from a source domain X to a target domain Y. Pix2pix [14] is the pioneering work that introduced CGAN [15] into image translation tasks. While Pix2pix can generate visually pleasing synthesized images, it relies on paired datasets for training. To address this limitation, subsequent methods like CycleGAN [16], DualGAN [17], DiscoGAN [18] and UNIT [19] are proposed, enabling image-to-image translation without the need for paired datasets. However, these methods cease to exhibit their advantages when handling more than two domains. Recognizing this deficiency, Choi et al. [10] propose StarGAN to learn the mapping relationships between multiple domains by leveraging the principle of CGAN. Notably, the discriminator in StarGAN is tasked with distinguishing the authenticity of samples as well as classifying which domain the sample originates from. AttGAN [20] refrains from imposing restrictions on the latent representation but instead enforces attribute classification constraints on the generated images to guarantee the accurate changes of the desired attributes. Although the above methods achieve multi-domain translation from two perspectives, there are still drawbacks. For example, StarGAN is constrained by its learning of a deterministic mapping for each domain, rendering it incapable of capturing the multi-modal characteristics of the data distribution. As a result, Choi et al. further optimize StarGAN and propose StarGAN v2 [11]. In contrast to prior work, this model introduces a mapping network and a style encoder as two key components. StarGAN v2 substitutes the original domain labels with domain-specific style codes, allowing it to convert images from a single domain into a variety of images across multiple

target domains. To achieve better multi-domain generation, Yang et al. propose W^{2}GAN [21] by introducing importance weighting and wavelet features to enhance the model's high-frequency awareness, while Ko et al. introduce SuperstarGAN [22], which replaces the auxiliary classifier with an independent one to improve style diversity and accuracy in image translation.

While StarGAN v2 has achieved significant successes in natural images, such as facial synthesis [23] and voice conversion [24], its performance in medical image analysis tasks is unsatisfactory. Specifically, it encounters three distinct challenges: (i) **Poor lesion generation capability**. In medical images, lesions refer to the areas that show signs of abnormalities or diseases, such as tumors and infections, which are very important for downstream medical analysis tasks. However, in StarGAN v2, the ability to generate lesions is unsatisfactory, sometimes the generated lesions are incomplete, and sometimes there is even no lesion area, especially when the reference image is a normal image, which lacks reservable lesion information during translation. (ii) **Insufficient tissue detail synthesis**. The incomplete rendering of tissue and organ intricacies within the image, to some extent hinders the performance of downstream analysis tasks. (iii) **High inter-modal similarity of results**. Each modal has its unique image features. For example, in T1 images, water and soft tissues exhibit low signal intensity (dark). While in Flair images, cerebrospinal fluid regions appear with low signal intensity (dark). These original style differences are not significant in the generated image. This prevents the full utilization of diverse information from multiple domains, thereby diminishing the supportive role of data augmentation in downstream tasks.

Considering the above background, to fully leverage the role of StarGAN v2 in medical image analysis tasks, especially when downstream tasks are segmentation tasks, we propose an Attentional Contrastive Learning-based Generative Adversarial Network (AttCL-GAN) to solve the three problems mentioned above. Therefore, to evaluate the performances of our proposed AttCL-GAN, StarGAN v2 [11], AttGAN [20], SuperstarGAN [22] and W^{2}GAN [21] are selected as baselines for multi-domain image-to-image translation in our experiments.

Contrastive learning. Contrastive learning proves to be a potent strategy for unsupervised representation learning [25–28]. It ensures the coherence of image representations across various augmentations by comparing positive pairs against negative ones. This approach has been investigated within a range of adversarial training scenarios [29–32]. For example, Zhao et al. [30] propose Cntr-GAN, which applies a contrastive loss to enforce regularization on the discriminator using two random augmented copies of both real and fake images. Kang et al. [33] propose ContraGAN for the task of conditional image generation. They leverage a novel conditional contrastive loss that is capable of learning relationships between both data and class. Cycle-consistency is a common method to solve disentanglement problem, which is the key issue in I2I translation. Nonetheless, cycle-consistency operates under the assumption that the connection between the two domains forms a bijection, which is a really strict condition. To address the above issue, CUT [34] proposes a method for image-to-image translation known as patch-based contrastive learning, which is achieved by leveraging positive pairs originating from the same locations within the input and output images. However, CUT employs features from random locations to enforce the constraint, a practice that may not be suitable, as some locations may not contain important information required for modal translation. Considering the above background, a Query-Selected Attention (QS-Attn) module is proposed [35], which involves the selection of pertinent anchor points, using them as queries to focus on and incorporate features from other locations. This process results in the creation of enhanced features suitable for contrastive learning.

Inspired by the application of contrastive learning in image generation and translation mentioned above, the first innovation of AttCL-GAN is to introduce an Attention-guided Contrastive Learning (ACL) module to improve the poor lesion generation capability of StarGAN v2. Consequently, to verify the effectiveness of introducing the new contrastive loss into StarGAN v2, experimental studies are conducted in this work. Due to the fact that most current contrastive learning techniques are predominantly utilized in cross-domain models, which are models that translate between two domains, like CycleGAN, to compare the application of contrastive learning in multi-domain models, we combine the contrastive loss module in CUT[34] with StarGAN v2 (denoted as StarGAN v2 + CUT), and the contrastive loss module in QS-Attn [35] with StarGAN v2 (denoted as StarGAN v2 + QS), using them as baselines for contrastive learning in our experiments.

Attention mechanisms. Drawing inspiration from the human attention mechanism [36], attention-based models have seen increasing adoption across a range of computer vision and machine learning tasks. In all of these tasks, attention enhances the performance by prompting the model to concentrate on the most pertinent areas of the input. As a result, many researchers have introduced the attention mechanisms into generative adversarial models. To effectively handle the holistic and shape-changing tasks, Kim et al. [37] propose a model named U-GAT-IT, which employs an attention module and AdaLIN for unsupervised image translation. However, the model's reliance on the auxiliary classifier for attention maps may introduce complexity and potential biases towards certain domain features. While proficient in translating low-level information, existing I2I methods fall short in capturing the high-level semantics of input images. To overcome this issue, AttentionGAN is proposed by Tang et al. [38], an unpaired I2I translation model that leverages attention mechanisms to identify and translate salient foreground objects while preserving the background. Nevertheless, the model's performance may be limited by the accuracy of the learned attention masks, which could miss subtle yet important details in complex images. As a result, Wang et al. [39]propose MSA-GAN, a multiscale attention-based GAN that incorporates Res2Net [40] and CBAM [41] modules to enhance structural preservation and detail fidelity in unsupervised I2I translation.

Motivated by the above mechanism and related works, to enhance the synthesis of tissue details, we have incorporated the attention mechanism into AttCL-GAN, which has two applications as follows. Firstly, in calculation of the contrastive loss, attention maps guide the selection of anchor points, which, combined with contrastive learning, becomes the first innovation of AttCL-GAN. Secondly, the same attention map is reused in the generator with AdaIN to better control image synthesis, named as Attentional AdaIN-based (AttAdaIN) generator, which is the second advancement of this article. Therefore, to validate the performance of StarGAN v2 with the introduction of attention mechanism, experiments are also conducted to compare AttCL-GAN with three state-of-the-art generative adversarial networks with attention: U-GAT-IT [37], AttentionGAN [38] and MSA-GAN [39], which are selected as three baselines for attention mechanism.

3. Methodology

Fig. 1 shows the overall structure of AttCL-GAN. Compared to the conventional StarGAN v2, AttCL-GAN mainly consists of three improvements: Attention-guided Contrastive Learning (ACL) module, Attentional AdaIN-based (AttAdaIN) generator and Style Code Diversity (SCD) loss. Next, in Section 3.1, we will first outline the overall framework of AttCL-GAN, then proceed to detail the implementation of three advancements in Sections 3.2–3.4, and summarize the objective function in Section 3.5. Additionally, the modality completion strategy will be described in Section 3.6.

3.1. The overall framework of AttCL-GAN

Let X and Y be the sets of images and possible domain labels, respectively. Given the source image $x_{src} \in X$ and its domain label $y_{src} \in Y$, as well as the reference image $x_{ref} \in X$ and its domain label $y_{ref} \in Y$, our

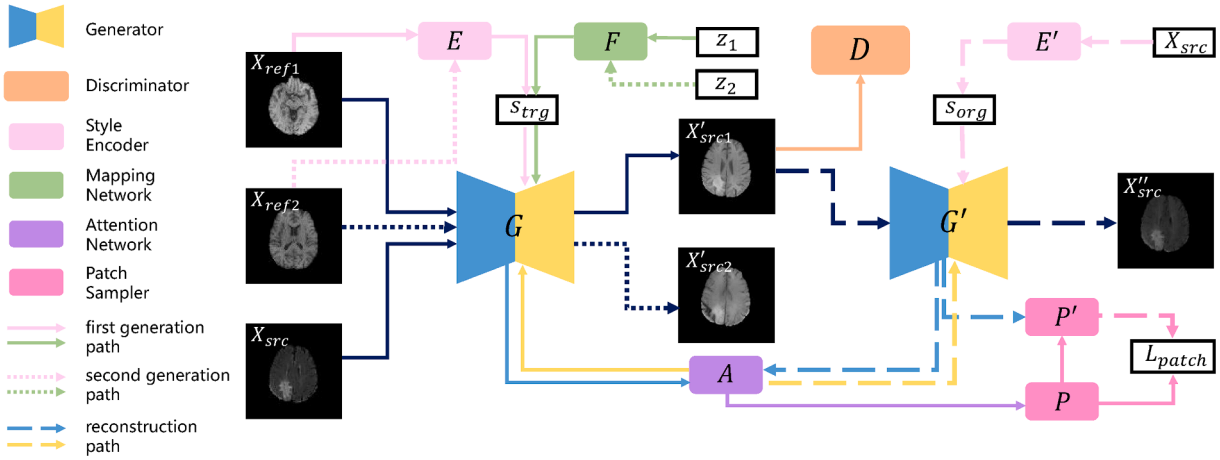


Fig. 1. Illustration of our proposed Attentional Contrastive Learning-based Generative Adversarial Network (AttCL-GAN). X_{ref}: Image of the reference modal. X_{src}: Image of the source modal. X'_{src}: Translated image of the reference modal. X''_{src}: Reconstructed image of the source modal. z: Random sampling of the latent vector. s_{trg}: Style code of the target/reference domain. s_{org}: Style code of the origin/source domain. The color of arrows indicates the source or destination of the data flow, representing which model component outputs the features or which component will receive them. Both the first generation path and the second generation path implement the image translation process from the source modal to the reference modal. They share the same generation principle, differing only in the input image and the latent code employed.

goal is to train a single generator G, that can generate images with the content of x_{src} and the domain style of x_{ref}, ultimately complementing the corresponding images of x_{src} in all modalities. Fig. 1 shows the overall structure of our model, which consists of six parts detailed below.

Generator. The generator, denoted as G, translates an input image x into an output image $G(x, s)$ that reflects a domain-specific style code s. The style codes are provided by either the mapping network or the style encoder.

Mapping network. Given a latent code z and a domain label y, the mapping network, denoted as F, generates the style code $s = F_y(z)$, where $F_y(\cdot)$ refers to an output of F corresponding to the domain label y. Diverse style codes are generated through random sampling of the latent vector $z \in Z$ and the domain label $y \in Y$.

Style encoder. Given an image x and its domain label y, the style encoder, denoted as E, extracts the style code $s = E_y(x)$ of x. Here, $E_y(\cdot)$ refers to an output of E corresponding to the domain label y. E is capable of generating diverse style codes using different reference images.

Attention network. Given a feature map F of certain NCE layer, the attention network, denoted as A, outputs the refined feature F_{refi}. A is implemented by the Convolutional Block Attention Module (CBAM), which first compresses the feature map in the spatial dimension to focus on important regions, and then compresses it in the channel dimension to focus on location information. The F_{refi} obtained in this way has two purposes: one is to guide the selection of anchor points in patch sampler, and the other is to combine it with AdaIN to feed spatial and channel attention information into the decoder of G to control better generation.

Patch sampler. Given the deep features F extracted by the encoder of G, when no positional information P_s is provided, the patch sampler, denoted as P, will leverage F_{refi} extracted by A, along with the cross entropy to calculate the importance matrix $I \in R^{HW \times HW}$. Consequently, the top i important patches are selected as queries in the translation, and their corresponding features F_q and positional information P_s are output separately. Unlike the above process, if P_s is provided, P will directly filter F at patch-level based on P_s, identify the corresponding keys in the generated image, and return the feature F_k contained in these keys. During the training phase, F_q and F_k are used to compute patch-level contrastive loss to strengthen constraints on G.

Discriminator. The discriminator, denoted as D, is a multi-task discriminator with multiple output branches. Each branch D_y is responsible for binary classification, determining whether an image is a real image x of corresponding domain label y or a fake image $G(x, s)$ generated by G.

3.2. Attention-guided contrastive learning module

The first improvement of AttCL-GAN is to introduce a Attention-guided Contrastive Learning (ACL) module into StarGAN v2 to alleviate the problem of poor lesion generation capability.

In contrastive learning, the fundamental idea involves associating a selected query with its relevant positive example while simultaneously distancing it from unrelated negatives. In this context, query represents an output, while the positive and negatives refer to corresponding and noncorresponding input respectively. The ultimate goal is to maximize the mutual information between input and output. Specifically, in the work of image translation, Taesung at al [34] note that content is shared not only across the whole images but also among corresponding patches in both input and output images. Therefore, patch-level contrastive learning has been proposed. Inspired by their work, in our model, the encoder of G, is also used to increase constraints on the output. While previous work [42,43] tends to favor the use of additional encoders to capture feature similarity between different domains, the intrinsic structure of G suggests that its encoding part is capable of performing feature extraction without adding model capacity. By extracting features from the input x and the output $G(x, s)$, the self-supervised patch-level contrastive loss is computed, with its specific formula shown in Eq. (1).

$$L_{patch} = -\log\left[\frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{i=1}^{N-1}\exp(q \cdot k^-/\tau)}\right] \tag{1}$$

where q refers to the feature patch from $G(x, s)$, k^+ and k^- respectively refer to one positive sample patch and $N-1$ negative sample patches from x. τ is a temperature hyper-parameter. As mentioned earlier, if q is a patch sampled from the generated image, then k^+ represents the patch from the real image that has the same position as q, while k^- stands for $N-1$ randomly selected patches from locations other than the above position in the real image. It is important to note that the gradient of L_{patch} only relates to q, not to k^+ and k^-, so G only learns the single direction of domain translation.

In CUT, the contrastive loss is computed with randomly selected q, k^+ and k^-, which may not be efficient in practical applications. Randomly

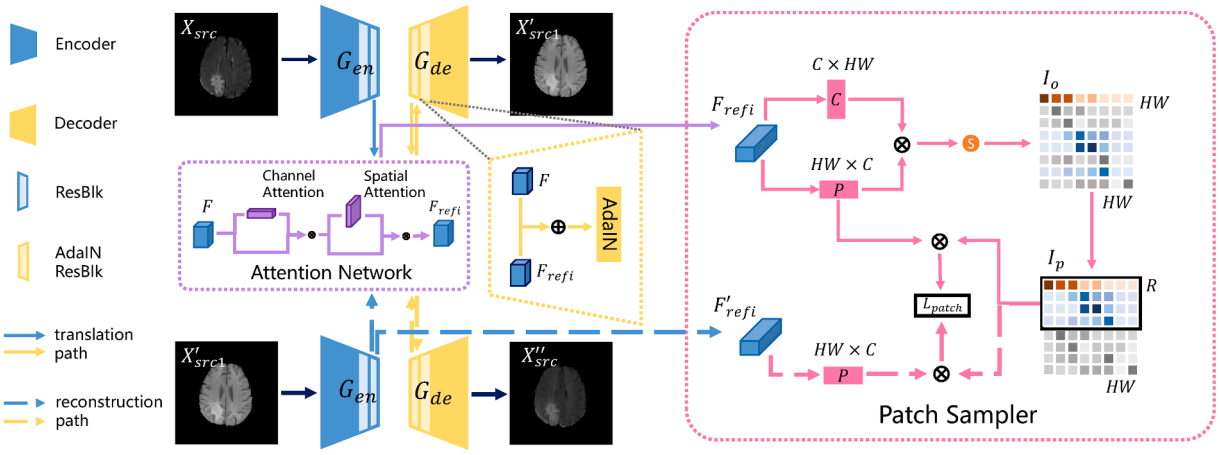


Fig. 2. The implementation diagram of the attention-guided contrastive learning (ACL) module and the attentional AdaIN-based (AttAdaIN) generator. X_{src}: Image of the source modal. X'_{src1}: One translated image of the reference modal. X''_{src}: Reconstructed image of the source modal. F: The extracted deep feature. F_{refi}: The refined feature. I_o: The overall importance matrix. I_p: The partial importance matrix. L_{patch}: Patch-level contrastive loss.

chosen patches might not necessarily contain crucial information that needs to be learned among different domains. For instance, in image translation tasks like horse-to-zebra, the body of animal is a key area that should be selected, while the irrelevant background behind can be ignored. Taking inspiration from the query selection idea in QS-Attn [35], we propose a contrastive learning module guided by attention information, which allows the model to select queries more efficiently. Fig. 2 shows the implementation details of the proposed ACL module.

Overall, the implementation of this module consists of three stages: In the first stage, given the deep feature $F \in R^{C\times H\times W}$ obtained from the encoder of G, CBAM first integrates spatial information by using average-pooling and max-pooling operations, generating two spatial information descriptors: F^c_{avg} and F^c_{max}. Then, two descriptors are fed into a shared network to generate a one-dimensional channel attention map $M_c \in R^{C\times 1\times 1}$, which is calculated as:

$$M_c(F) = \sigma(MLP(\text{AvgPool}(F)) + MLP(\text{MaxPool}(F))) \tag{2}$$

where σ denotes the sigmoid function and MLP represents the multi-layer perceptron with one hidden layer. Secondly, using the same pooling operations, CBAM generate two two-dimensional descriptors: F^s_{avg} and F^s_{max}. They are first concatenated in the first dimension, and then convolved using a standard convolutional layer, resulting in a two-dimensional spatial attention map $M_s \in R^{H\times W}$, which is computed as:

$$M_s(F) = \sigma\left(f^{7\times 7}([\text{Avg Pool}(F); \text{MaxPool}(F)])\right) \tag{3}$$

where $f^{7\times 7}$ represents a convolution operation with the filter size of 7×7. And the overall attention process can be summarized as:

$$F' = M_c(F) \otimes F \tag{4}$$

$$F'' = M_s(F') \otimes F' \tag{5}$$

where $\otimes$ denotes element-wise multiplication.

In the second stage, based on the refined feature obtained in the first stage, a term named entropy E_n is utilized to measure the importance of each position in F, which is first adopted in [35] . Specifically, given $F_{refi} \in R^{C\times H\times W}$ extracted from the source image x_{src}, which is refined by the CBAM and has the same size as F_{src}, we first reshape it into a two-dimensional matrix $P_{src} \in R^{HW\times C}$ and then multiply it by its transposition $C_{src} \in R^{C\times HW}$. Then, we operate on each row of the product result using the softmax function to obtain the overall importance matrix $I_o \in R^{HW\times HW}$. Finally, we use Eq. (6) to calculate the entropy of each row in I_o, which measures the importance of patches at different positions.

$$E_n(i) = -\sum_{j=1}^{HW} I_o(i,j) \log I_o(i,j) \tag{6}$$

where i and j represent the index of row and column in I_o. The meaning of $E_n(i)$ is that in the i-th row, how many patches of other positions are similar to the i-th patch. Therefore, the smaller $E_n(i)$ is, the more unique the i-th patch is. Thus, it should be selected as a query to calculate contrastive loss. To select the most important patches as queries, we sort the rows of I_o according to the ascending order of $E_n(i)$ and choose the smallest R rows to form the partial importance matrix I_p. It is worth noting that both importance matrix I_o and I_p is generated from the input x and has no relation to the output $G(x, s)$.

In the third stage, we feed the generated image $G(x, s)$ into G, similarly obtain its deep feature F using the last layer of encoder, and pass it through the CBAM. This involves the same attention extraction procedure as in the first stage, yielding the refined feature F'_{refi}. Then, F'_{refi} is reshaped into a two-dimensional matrix $P_{ref} \in R^{HW\times C}$. We perform element-wise multiplication of P_{src} and P_{ref} with the partial important matrix I_p, respectively, and calculate the L_{patch} using the obtained results based on Eq. (1).

3.3. Attentional AdaIN-Based generator

The second improvement of AttCL-GAN is to utilize an Attentional AdaIN-based (AttAdaIN) generator to address the issue of insufficient tissue detail synthesis.

Vanilla AdaIN [44] takes content information x and style information y as input, adjusting the channel-wise mean and variance of x to maintain the same as those of y. Unlike BN, IN, and CIN, AdaIN does not have affine parameters that can be learned by the network. Instead, it adaptively computes the affine parameters from the style information y according to the following formula.

$$\text{AdaIN}(x, y) = \sigma(y)\left(\frac{x - \mu(x)}{\sigma(x)}\right) + \mu(y) \tag{7}$$

Specifically, it scales the content information x by $\sigma(y)$ and shifts x by $\mu(y)$, where $\sigma(y)$ is the standard deviation of y, and $\mu(y)$ is the mean of y. This is a common normalization method used in style transfer.

However, since the mean and variance of features are calculated globally, local details and some pixel-level patterns are mostly ignored [45], leading to the degradation of local style transfer effects in the output. This is the problem of insufficient tissue detail synthesis mentioned earlier. Based on such issue, previous efforts to improve normalization have increased overall complexity by incorporating additional input information, which is often derived through individual feature extraction network [46,47]. This places a considerable training burden on the already large-capacity generative adversarial networks, and it may even

hinder the final generation of images in the inference stage due to the lack of annotated information. Consequently, we consider reutilizing the F_{refi} obtained from the first stage of ACL module to improve the normalization operation within the decoder of G.

Specifically, since our generation goal is to blend the content from x_{src} and the style from x_{ref}, attention extraction in both spatial and channel dimensions is more crucial for x_{src} compared to x_{ref}. Therefore, we perform element-wise addition on F_{refi} and the original F_{src} at the same layer to obtain $\tilde{F}_{src}$. After normalizing $\tilde{F}_{src}$, following the concept of AdaIN, we then align the mean and variance between the content input $\tilde{F}_{src}$ and the style input s. The calculation formula for the entire process is as follows.

$$\tilde{F}_{src} = F_{refi} \oplus F_{src} \tag{8}$$

$$\mathrm{AdaIN}(\tilde{F}_{src}, s) = \sigma(s)\left(\frac{\tilde{F}_{src} - \mu(\tilde{F}_{src})}{\sigma(\tilde{F}_{src})}\right) + \mu(s) \tag{9}$$

where $\oplus$ denotes element-wise addition. Fig. 2 shows the above implementation of the proposed AttAdaIN generator.

3.4. Style code diversity loss

A good diversity loss is highly valuable for generative adversarial models. On one hand, it can alleviate the problem of mode collapse. On the other hand, it can significantly enhance the diversity of generated results. Existing methods mostly rely on the idea of utilizing two different generated images, I_1 and I_2, to learn distinct features between modalities by maximizing the distance between them. Such as the commonly used mode-seeking loss [48], specifically calculated using the following equation.

$$L_{ms} = \mathbb{E}\left[\frac{d_x(\boldsymbol{I}_1, \boldsymbol{I}_2)}{d_z(\boldsymbol{z}_1, \boldsymbol{z}_2)}\right] = \mathbb{E}_{\boldsymbol{x}, \boldsymbol{z}_1, \boldsymbol{z}_2}\left[\frac{\|G(\boldsymbol{x}, \boldsymbol{z}_1) - G(\boldsymbol{x}, \boldsymbol{z}_2)\|_1}{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_1}\right] \tag{10}$$

where $d(\cdot)$ represents the distance calculated by the L_1 norm $\|\cdot\|_1$, z_i represents the noise vectors, and $G(x, z_i)$ represents the generated images. However, when applying mode-seeking loss in practical applications, Choi et al. [11] found that minor variations in the denominator could significantly increase in the overall loss, causing instability in the training process. Therefore, in StarGAN v2, the denominator $\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_1$ is directly removed, and the generator is constrained solely by maximizing the distance between the two generated images. This loss is named diversity-sensitive loss. While the diversity-sensitive loss has achieved certain effectiveness, it ultimately lacks constraints in the latent space due to the direct omission of the denominator.

Therefore, as style code is the numerical representation of style information in the latent space, we consider using it to impose constraints on G. In the process of extracting style information using F or E, the original output is a set of codes $s_{all} \in R^{N \times M}$, where the first dimension N is the total number of possible domains, and the second dimension M is the number of digits in each style code, a hyper-parameter. For the input image x, s_{all} includes not only the style code for the modal corresponding to the label y, but also the latent style representation $s_i \in R^M$ in the remaining non-corresponding modalities. To enhance the style differences between different domains, we propose calculating the average Euclidean distance between the style codes corresponding to the modal y and other modalities, named as style code diversity loss. The specific calculation formula is as follows.

$$L_{scd} = \frac{1}{N}\sum_{i=0}^{N-1}\left\|s_y - s_i\right\|_2 \tag{11}$$

where s_y denotes the style code corresponding to modal y, and s_i denotes the style code for any other possible modalities, both in the context of the input image x, and $\|\cdot\|_2$ denotes the L_2 norm, which is also called the Euclidean distance. By maximizing this loss, we remedy the problem of high inter-modal similarity of results. At the same time, this design neither introduces new network modules nor adds additional training burden.

3.5. Objective functions

Given the source and reference images: x_{src} and x_{ref}, along with their corresponding domain labels: y_{src} and y_{ref}, we use the following objectives to train the model.

Adversarial loss. G takes the image x_{src} and the style code s as inputs, continuously learns through adversarial loss, and ultimately produces a desirable output image $G(x_{src}, s)$. The calculation formula for this loss is as follows:

$$L_{adv} = \mathbb{E}_{x_{src}, y_{src}}[log D_{y_{src}}(x_{src})] + \mathbb{E}_{x_{src}, y_{ref}, z}[log(1 - D_{y_{ref}}(G(x_{src}, s)))] \tag{12}$$

where $D_y(\cdot)$ refers to the output of D specific to domain y. During this period, the goal of F is to generate style code that better align with the characteristic of y_{ref}, and G is dedicated to producing a more realistic output $G(x, s)$ corresponding to y_{ref} through s.

Style reconstruction loss. Constraining G solely with adversarial loss is far from sufficient. Therefore, we apply the style reconstruction loss, which is calculated as follows:

$$L_{sty} = \mathbb{E}_{x_{src}, y_{ref}, z}[\| s - E_{y_{ref}}(G(x_{src}, s)) \|_1] \tag{13}$$

Unlike previous methods that require multiple encoders to map images to latent codes for different target domains, we train a single style encoder E to achieve multi-domain outputs. During the testing phase, E assists G in style transfer for input image x_{src}, ensuring that the generated results carry the style characteristic of x_{ref}.

Diversity sensitive loss. This is the original regularization imposed on G to enhance output diversity in StarGAN v2. It performs mathematical calculations from the perspective of generated images.

$$L_{ds} = \mathbb{E}_{x_{src}, y_{ref}, z_1, z_2}[\| G(x_{src}, s_1) - G(x_{src}, s_2) \|_1] \tag{14}$$

By maximizing this regularization, G can better explore the latent space to discover more significant style features, ultimately contributing to diverse output results.

Cycle consistency loss. The above two losses ensure the correct transfer of style, but the content from x_{src} also needs to be considered. Therefore, we employ the classic cycle consistency loss to maintain domain-independent information during generation.

$$L_{cyc} = \mathbb{E}_{x_{src}, y_{src}, y_{ref}, z}[\| x_{src} - G(G(x_{src}, s), \tilde{s}) \|_1] \tag{15}$$

where $\tilde{s} = E_{y_{src}}(x')$ is obtained given the generated image x'. We feed the estimated $\tilde{s}$ into G to reconstruct x_{src}. Through this process, G learns to efficiently change the style to y_{ref} while preserving the content of x_{src}.

Therefore, combined with the patch-level contrastive loss discussed in Section 3.2 and the style code diversity loss proposed in Section 3.4, our full objective function can be summarized as follows.

$$\min_{G,F,E,A,P}\max_{D} L_{adv} + \lambda_{sty}L_{sty} + \lambda_{cyc}L_{cyc} + \lambda_{patch}L_{patch} - \lambda_{ds}L_{ds} - \lambda_{scd}L_{scd} \tag{16}$$

where λ_{sty}, λ_{ds}, λ_{cyc}, λ_{patch}, λ_{scd} are hyper-parameters for each respective term.

3.6. Modality completion strategy

To complement the missing modal images using our multi-domain generative model, we further propose a modality completion strategy. Firstly, from an overall perspective, it is necessary to store the original modal missing dataset and the complemented full dataset in two different paths. The purpose of this operation is to ensure that existing multi-modal images are not overwritten by newly generated images in the subsequent data imputation process. Secondly, it is essential to store names of all possible modalities in a variable called *modal_list*, in the order of decreasing data quantity for each modality. For the input image x_{src}, this variable determines the order of modal translation. As

shown in Table 2, the FID and IPIPS for translations from the modality with the largest quantity (e.g., Flair in ratio1) to other modalities (e.g., T1, T1ce, T2 in ratio1) are lower compared to other directions, which can be theoretically explained. Modalities with a larger number of images contain more diverse and detailed information, aiding the model in accurately capturing the mapping relationships. This helps in learning and preserving useful details and modal-specific features, making the generated images more consistent with the distribution of real medical images, and yield lower values of generation metrics. Therefore, we prioritize translations from the modality with the largest number of images to other modalities. We employ a path-based check to prevent subsequently generated images from other modalities from overwriting existing ones, acheving modality completion of all samples. In addition, unlike the training phase, during the sampling process, each modality retains only one image in the reference image set to expedite the efficiency of completion task. The detailed strategy is shown in Algorithm 1.

Algorithm 1 Modality completion strategy.

Input: x_{src}: a source image, x_{ref}: collection of reference images, $slice\_len$: number of all source images, y_{ref}: modality of reference images, $path_{src}$: path of the source image, $modal\_list$: list of all possible modalities, s_{ref}: collection of style codes, x': a generated image, $modal\_src$: modality of x_{src}, suf: suffix of image name, $path_{ori}$: path of the original dataset, $path_{cpl}$: path of the complementing dataset.

Obtain x_{ref}.
for $i = 0$ **to** $slice\_len - 1$ **do**
  Obtain the i-th source image x_{src} and its path $path_{src}$.
  Given x_{ref} and y_{ref}, use E to obtain s_{ref}.
  for $j = 0$ **to** $modal\_number - 1$ **do**
    Obtain the style code $s_{ref}(j)$ belonging to the j-th modality.
    Given x_{src} and $s_{ref}(j)$, use G to generate x'_{src} with the style of the j-th modality.
    Based on $path_{src}$, extract $modal\_src$ and suf in the format of "patient number_slice number".
    Concatenate $path_{ori}$, $s_{ref}(j)$, and suf to obtain $corr\_path_{ori}$.
    if x_{src} has a paired image corresponding to $modal\_list(j)$ in the original location $corr\_path_{ori}$. **then**
      Do nothing and enter the next FOR loop.
    else
      Concatenate $path_{cpl}$, $s_{ref}(j)$, and suf to obtain $corr\_path_{cpl}$.
      if x_{src} has a paired image corresponding to $modal\_list(j)$ in the complementing location $corr\_path_{cpl}$. **then**
        Do nothing and enter the next FOR loop.
      else
        Store x'_{src} in the new location $path_{cpl}$ corresponding to $modal\_list(j)$.
      end if
    end if
  end for
end for

4. Datasets and experiments

4.1. Datasets

To evaluate the performances of our proposed AttCL-GAN in medical image generation and segmentation tasks, we conduct extensive experiments on a multi-modal brain tumor datasets (BraTS 2021[1]). The BraTS 2021 dataset, provided by Baid et al. [49], is a public real-world collection of multi-modal magnetic resonance imaging (MRI) data with the primary goal of glioma tumor segmentation. Each sample in the dataset consists of four modalities: fluid attenuation inversion recovery (Flair), T1 weighting (T1), T1-weighted contrast-enhanced (T1ce), and T2 weighting (T2). And the ground truth masks are labeled by expert board-certified neuroradiologists.

Firstly, in this dataset, each patient has complete images from all four modalities, making it an ideal balanced multi-modal dataset and an excellent resource for constructing modality missing scenarios. Therefore, to assess the performance of AttCL-GAN in multi-modal data imputation tasks, based on BraTS 2021, we designed the following three scenarios: (i) **Scenario 1**: Three modalities have missing data. The first third of patients have data from only two modalities, with T1ce and T2 modal images missing. The middle third of patients have data from two modalities, with T1 and T2 modal images missing. Finally, the last third of patients have data from two modalities, with T1 and T1ce modal images missing. Consequently, the proportion of patients for four modalities is Flair:T1:T1ce:T2 = 1:1/3:1/3:1/3. This scenario is labeled as ratio1. (ii) **Scenario 2**: Two modalities have missing data. The first half of patients have data from three modalities, with T2 modal image missing. Meanwhile, the second half of patients have data from three modalities as well, with T1ce modal image missing. Therefore, the proportion of patients for four modalities is Flair:T1:T1ce:T2 = 1:1:1/2:1/2. We mark this scenario as ratio2. (iii) **Scenario 3**: Only one modality has missing data. The first half of patients, representing 1/2 of the dataset, have complete information with all modalities. The remaining half of patients have data from 3 modalities, with T2 modal image missing. The proportion of patients for four modalities is Flair:T1:T1ce:T2 = 1:1:1:1/2, and we label this scenario as ratio3.

Although the three scenarios represent different degrees of modality incompleteness, they are essentially designed by setting one or two modalities with larger sample sizes to 1, while the remaining modalities are divided according to their relative quantities. As a result, all possible data missing situations can be transformed into the three designed cases, which are sufficient and reasonable for validating the data imputation capability of the proposed model.

Secondly, to evaluate the model's performance on small dataset approaching real-world scenarios, based on the three cases of modal absence, we further divided a small sample dataset from the original dataset and name it BraTS$_s$. Detailed information regarding this partition is provided in the following text. The statistical information about the final constructed datasets are shown in Table 1.

BraTS: The dataset consists of 243 cases. We allocated 70 % of them as training data, 20 % as testing data and 10 % as validation data. Each patient sample includes four 3D volumes representing different modalities, along with corresponding ground truth for brain tumors. We performed preprocessing by slicing the 3D MRI volumes, each initially sized at $240 \times 240 \times 155$, into 155 individual 2D slices. Given that the image quality in the first 30 slices and the last 30 slices is relatively poor, we opted to focus on the middle slices between 30 to 125.

BraTS$_s$: A subset of BraTS, denoted as BraTS$_s$, is used to verify the effectiveness of our data augmentation approach with a restricted data volume. The training set of BraTS$_s$ is one eighth of that of BraTS. To conduct a fair assessment of the model performance across different scales, we kept the validation and testing set consistent with BraTS.

4.2. Baselines

The main purpose of AttCL-GAN is to complement datasets with modal data absence through multi-domain image-to-image translation. These augmented datasets are then utilized in the training of downstream segmentation models, ultimately improving the accuracy of segmentation results. Therefore, to verify the performances of the proposed AttCL-GAN, we compare the model with i) multi-domain image-to-image translation methods [11,20–22] and ii) GAN-based data augmentation methods with contrastive learning [34,35]. Besides, AttCL-GAN is further compared with iii) GAN-based data augmentation methods with attention mechanism [37–39].

[1] https://www.med.upenn.edu/cbica/brats2021/

Table 1
Statistical description of the three cases of modal absence. Datasets are divided by patients rather than the number of slices, with each patient having 94 slices.

Dataset		Modal proportion	Training set					Validation set					Testing set				
Modality		Flair:T1:T1ce:T2	Flair	T1	T1ce	T2	total	Flair	T1	T1ce	T2	total	Flair	T1	T1ce	T2	total
ratio1	BraTS	1:1/3:1/3:1/3	170	57	57	56	170	23	8	8	7	23	50	17	17	16	50
	$BraTS_s$		21	7	7	7	21										
ratio2	BraTS	1:1:1/2:1/2	170	170	85	85	170	23	23	12	11	23	50	50	25	25	50
	$BraTS_s$		21	21	11	10	21										
ratio3	BraTS	1:1:1:1/2	170	170	170	85	170	23	23	23	12	23	50	50	50	25	50
	$BraTS_s$		21	21	21	11	21										

i) **Multi-domain image-to-image translation methods**: StarGAN v2 [11], serving as the backbone of AttCL-GAN, is naturally chosen as one of the baselines. Similar to StarGAN [10], AttGAN [20] is a learning-based method for editing multiple facial attributes. In our experiments, it is trained on medical images for multi-modal data augmentation. In addition, W^2GAN [21] and SuperstarGAN [22] are novel multi-domain generation methods in recent years. Including them as comparison methods helps enhance the cutting-edge nature and representativeness of the experiments.

ii) **GAN-based data augmentation methods with contrastive learning**: Most current contrastive learning techniques are used in cross-domain models like CycleGAN. To compare their application in multi-domain models, we integrate the contrastive loss module from CUT [34] into StarGAN v2 (denoted as StarGAN v2 + CUT) and the contrastive loss module from QS-Attn [35] into StarGAN v2 (denoted as StarGAN v2 + QS).

iii) **GAN-based data augmentation methods with attention mechanism**: U-GAT-IT [37], AttentionGAN [38] and MSA-GAN [39] are innovative I2I translation models using different attention mechanisms to generate realistic images. However, both are limited to cross-domain generation. Therefore, to compare the two methods with our model, we separately utilized images from different modalities, and trained each model three times to achieve translations between four domains, accomplishing the modal data imputation.

4.3. Evaluation metrics

In order to show the overall workflow of AttCL-GAN more clearly, we divided the method into two parts: upstream generation and downstream segmentation. Firstly, the performance of upstream generation is evaluated using two metrics. Fréchet inception distance (FID) is used to measure the similarity between the distribution of generated images and real images. The formula for FID is as follows:

$$FID = \left\|\mu_r - \mu_g\right\|_2^2 + Tr(\sum_r + \sum_g - 2(\sum_r \sum_g)^{\frac{1}{2}}) \tag{17}$$

where μ is the mean of the image feature, $\sum$ is the covariance matrix of the image feature. Learned perceptual image patch similarity (LPIPS) measures the quality and structural differences between images by comparing their perceptual similarity. Instead of a simple mathematical equation, LPIPS is learned through training a perceptual similarity network, so an exact formula is not available.

Secondly, to validate the improvement in downstream segmentation performance with the augmented datasets, four widely used evaluation metrics are employed. Dice similarity coefficient (Dice) refers to the similarity between the predicted segmentation region and the real segmentation region. Sensitivity (Sens) measures the proportion of actual positive samples correctly identified by the model. Intersection over union (IoU) is used to measure the ratio of the intersection of two sets relative to their union, providing a way to quantify the degree of overlap. Their calculation formulas are as follows:

$$Dice = \frac{2 * TP}{2 * TP + FP + FN} \tag{18}$$

$$Sens = \frac{TP}{TP + FN} \tag{19}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{20}$$

where TP, FP, and FN are the number of true positive points, false positive points, and false negative points, respectively.

Hausdorff Distance (HD) measures the maximum deviation between predicted and real segmentation boundaries. HD95, the 95th percentile of HD, enhances robustness by reducing sensitivity to outliers in segmentation evaluation as follows:

$$HD95 = max_{k95\%}[\sum_{p \in P} min_{g \in G} d(p, g), \sum_{g \in G} min_{p \in P} d(g, p)] \tag{21}$$

where G represents the real label, and P represents the segmentation result.

4.4. Implementation settings

All models are implemented using the PyTorch framework[2] and run on a server equipped with 8 Nvidia GeForce 2080 Ti GPUs. Each graphics card has 11019M of memory, and the server features an Intel(R) Xeon(R) Silver 4110 CPU with a clock speed of 2.10GHz and 16GB of RAM. Regarding software configuration, the server's CUDA version is 10.2, and all code is implemented in Python language. Key Python libraries utilized in the experiments include Numpy, PIL, Nibabel, visdom, and torchvision. The Adam algorithm is adopted for optimizing the networks.

For generating tasks, the learning rate is initialized to 1e-4 for the gernerator, style encoder, discriminator, patch sampler and attention network. The learning rate for the mapping network is set to 1e-6. The decay rates for the first and second moments of the Adam optimizer are set to 0.0 and 0.99 respectively. Additionally, a weight decay of 1e-4 is applied, and during sampling, 10 generated images are produced per domain. For segmentation tasks, the learning rate for the model optimization using Stochastic Gradient Descent (SGD) is set to 1e-3, and the momentum term is set to 0.9. Additionally, a weight decay term of 1e-4 is applied to regularize the weights during optimization.

Overall, key hyperparameters were determined through grid search. Specifically, the ranges of learning rate were set as follows: [1e-2, 1e-3, 1e-4, 1e-5] for the generator, style encoder, discriminator, patch sampler and attention network; [1e-4, 1e-5, 1e-6, 1e-7] for the mapping network; and [1e-2, 1e-3, 1e-4] for the segmentation network. Other specific hyperparameters were selected through empirical tuning. The batch size was chosen based on the capacity and performance of the GPUs used, set to the maximum value supported by the hardware.

[2] https://pytorch.org/

Table 2
Comparison with the state-of-the-art GAN-based data augmentation methods in the upstream generation stage. The mean row is calculated by averaging the six FID values obtained from the same model across three translation tasks, each performed in two directions.

Metric			BraTS				$BraTS_s$			
Direction			FID		LPIPS		FID		LPIPS	
Method			A-to-B	B-to-A	A-to-B	B-to-A	A-to-B	B-to-A	A-to-B	B-to-A
ratio1	Flair-T1	StarGAN v2	56.1939	62.9818	0.0413	0.0675	81.8387	82.7089	0.0535	0.0791
		AttGAN	61.3541	66.2227	0.0442	0.0691	87.7406	85.5803	0.0562	0.0801
		StarGAN v2 + CUT	47.6417	60.6704	0.0381	0.0641	73.6545	79.8813	0.0511	0.0767
		StarGAN v2 + QS	41.1223	54.9565	0.0327	0.0560	68.7305	72.8069	0.0472	0.0673
		U-GAT-IT	46.4584	59.4931	0.0367	0.0624	73.0533	79.0751	0.0491	0.0751
		AttentionGAN	42.2165	56.9814	0.0334	0.0573	70.2681	76.5682	0.0488	0.0696
		MSA-GAN	43.7842	55.8173	0.0328	0.0566	70.0525	74.9157	0.0473	0.0709
		SuperstarGAN	40.5951	53.1580	0.0315	0.0558	66.7291	71.2914	0.0446	0.0668
		W^2GAN	40.9676	54.3704	0.0319	0.0570	68.1152	72.4906	0.0458	0.0682
		AttCL-GAN (Ours)	**38.6402**	**51.3290**	**0.0302**	**0.0542**	**64.0055**	**69.5564**	**0.0419**	**0.0651**
	Flair-T1ce	StarGAN v2	79.4471	94.0625	0.0563	0.0748	96.4698	108.7358	0.0641	0.0866
		AttGAN	80.6428	96.3247	0.0586	0.0760	98.2923	111.2441	0.0692	0.0895
		StarGAN v2 + CUT	73.5382	84.8473	0.0514	0.0691	89.8102	99.7035	0.0598	0.0837
		StarGAN v2 + QS	66.6110	79.0827	0.0465	0.0675	83.2155	93.5336	0.0558	0.0791
		U-GAT-IT	69.3852	82.3403	0.0501	0.0699	87.2084	97.9322	0.0577	0.0820
		AttentionGAN	67.2481	79.1005	0.0492	0.0673	84.3641	94.9114	0.0570	0.0787
		MSA-GAN	67.5974	78.2944	0.0482	0.0682	83.5986	93.8079	0.0563	0.0805
		SuperstarGAN	64.5185	77.0849	0.0460	0.0671	79.1508	91.2667	0.0544	0.0781
		W^2GAN	65.3406	76.1856	0.0473	0.0684	81.4870	91.5924	0.0539	0.0773
		AttCL-GAN (Ours)	**62.4443**	**75.0615**	**0.0456**	**0.0653**	**78.5484**	**89.2963**	**0.0526**	**0.0764**
	Flair-T2	StarGAN v2	41.2260	52.6312	0.0408	0.0571	69.1525	73.0769	0.0497	0.0685
		AttGAN	44.6732	56.4701	0.0452	0.0580	72.0125	79.2204	0.0521	0.0717
		StarGAN v2 + CUT	33.9296	51.0198	0.0366	0.0530	65.5104	68.7323	0.0443	0.0660
		StarGAN v2 + QS	26.2479	44.4498	0.0299	0.0488	55.9914	61.0726	0.0372	0.0598
		U-GAT-IT	34.6513	45.5144	0.0349	0.0514	62.2140	64.0441	0.0460	0.0672
		AttentionGAN	28.0041	46.0771	0.0313	0.0499	56.7051	67.8813	0.0382	0.0631
		MSA-GAN	26.9818	45.8048	0.0325	0.0512	55.4168	63.7315	0.0405	0.0623
		SuperstarGAN	25.4921	42.9470	0.0292	0.0483	52.9246	59.1351	0.0364	0.0602
		W^2GAN	25.8134	43.8792	0.0304	0.0496	53.1889	60.9804	0.0378	0.0594
		AttCL-GAN (Ours)	**23.8955**	**40.6102**	**0.0287**	**0.0479**	**49.5210**	**57.1813**	**0.0351**	**0.0587**
	mean	StarGAN v2	64.4238		0.0563		85.3304		0.0669	
		AttGAN	67.6146		0.0585		89.0150		0.0698	
		StarGAN v2 + CUT	58.6078		0.0521		79.5487		0.0636	
		StarGAN v2 + QS	52.0784		0.0469		72.5584		0.0577	
		U-GAT-IT	56.3071		0.0509		77.2545		0.0629	
		AttentionGAN	53.2713		0.0481		75.1164		0.0592	
		MSA-GAN	53.0467		0.0483		73.5872		0.0596	
		SuperstarGAN	50.6326		0.0463		70.0830		0.0568	
		W^2GAN	51.0928		0.0474		71.3091		0.0571	
		AttCL-GAN (Ours)	**48.6635**		**0.0453**		**68.0182**		**0.0550**	

5. Results

5.1. Main results

Based on the discussions in the above four sections, in the main experiment, we evaluate the performance of AttCL-GAN through two stages: the upstream generation stage and the downstream segmentation stage. The generation results of AttCL-GAN and nine state-of-the-art GAN-based data augmentation methods on BraTS and $BraTS_s$ are shown in Table 2, and the visualized results for 4 slices are presented in Fig. 3. After the upstream generation, the complete datasets, obtained through modal data imputation by these methods, are used for the segmentation task. The results of this stage are presented in Table 3.

5.1.1. Upstream generation stage

As shown in Table 2, based on the two generation metrics, it can be seen that AttCL-GAN achieves the best generation performance among the seven models.

From the perspective of model performance, we can find the following conclusions:

i) StarGAN v2 + CUT and StarGAN v2 + QS are generally better than StarGAN v2 and AttGAN in all translation directions in terms of all metrics. This proves that the introduction of contrastive learning indeed enhances the learning capability of the generative models, making the feature distribution of generated images closer to real images. Additionally, despite both being contrastive learning-based models, StarGAN v2 + QS exhibits lower metrics in all directions compared to StarGAN v2 + CUT. This experimental result supports the notion that selecting anchor points with emphasis, as opposed to random selection, is more effective in playing the role of contrastive learning.

ii) U-GAT-IT, AttentionGAN and MSA-GAN consistently outperform the two basic multi-domain translation models. This is attributed to the introduction of attention mechanisms, which enhance the models' ability to learn key lesion regions and tissue details, resulting in more medically realistic generated images. This validates the correctness of incorporating attention mechanisms into generative models. Moreover, there is no significant performance difference between the models with attention mechanisms and those with contrastive learning. They represent two different approaches to optimizing generative models.

iii) W^2GAN and SuperstarGAN achieve the third and second best generation results, respectively. W^2GAN's integration of an importance weighting mechanism and wavelet feature guidance helps maintain

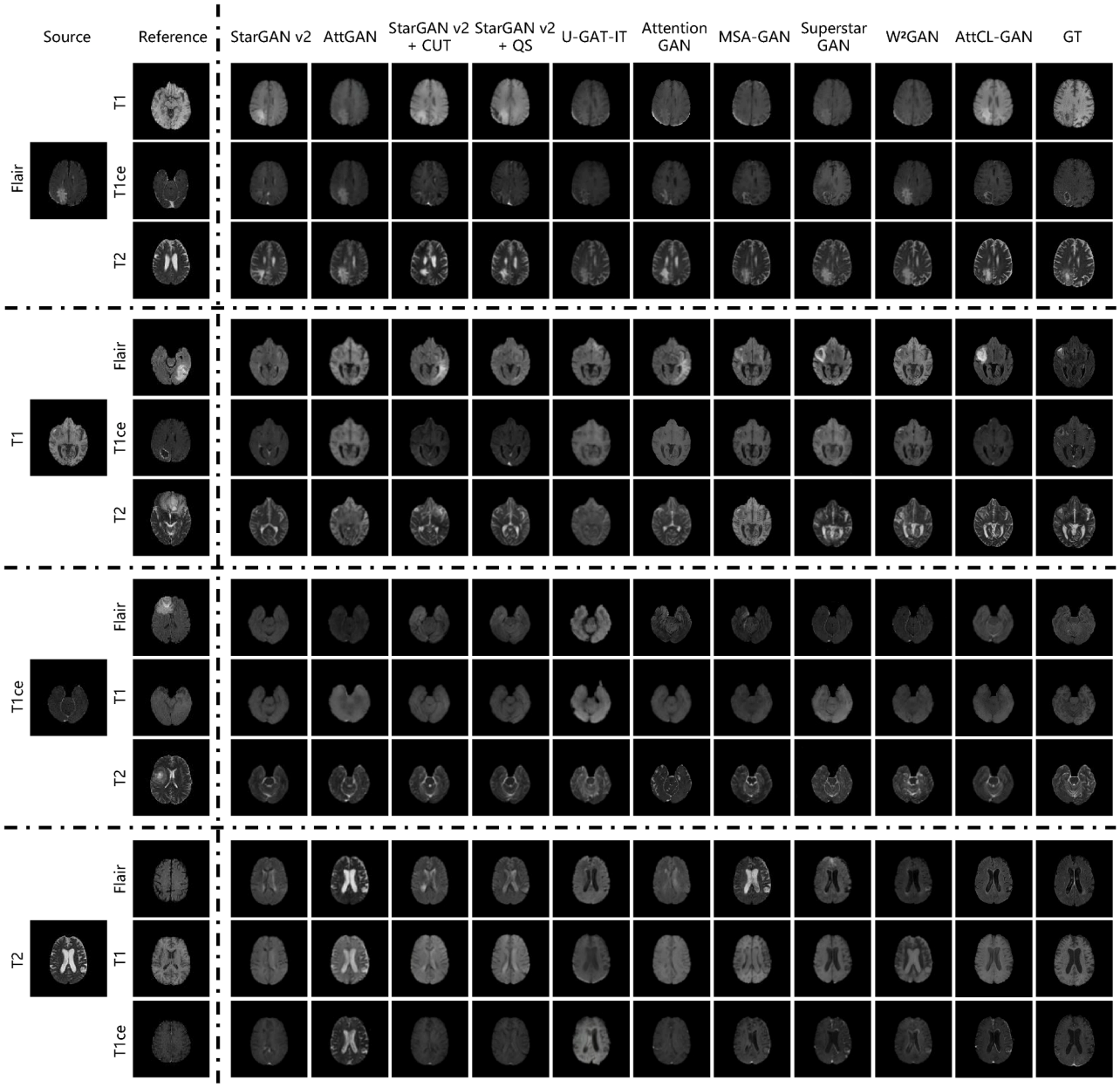


Fig. 3. The multi-modal images synthesized by ten methods in the four modal translations, along with the ground truth.

stable structural integrity and high-frequency details in generated images. SuperstarGAN enhances the model's capability to capture complex semantic information by replacing the original auxiliary classifier with an independent classifier and incorporating data augmentation techniques, resulting in target images with clear textures and consistent styles.

iv) We find that AttCL-GAN outperforms all the baselines in all translations. This demonstrates that by integrating attention mechanisms and contrastive learning, along with the addition of an extra diversity loss, we have achieved better generation of multi-domain translation models in medical images.

Additionally, from the perspective of modality translation, we observed the other two points:

v) Almost all metrics for forward translations (A-to-B) are superior to those for reverse translations (B-to-A). This is because, in the three types of translations, class A has more images compared to class B. This result aligns with the principle that more training data leads to richer input information, enabling the model to learn more universal and generalizable features and patterns.

vi) Among the three types of translations, Flair to T2 performs best, followed by Flair to T1, with Flair to T1ce performing the least well. This can be explained by medical imaging characteristics. Flair and T2 both highlight high water content tissues, making translation easier due to their similarity. Flair and T1, while providing structural information, differ in contrast and emphasized tissues, making translation moderately challenging. T1ce, with vessels and lesion features after contrast agent injection, differs significantly in imaging purpose and characteristics from Flair, making their conversion the most difficult and complex.

From the images synthesized by different methods in Fig. 3, we can draw several notable conclusions:

i) Compared to the other seven baselines, both StarGAN v2 and AttGAN fail to ensure the preservation and generation of lesions after modal

Table 3
Comparison with the state-of-the-art GAN-based data augmentation methods in the downstream segmentation stage. w/o DA represents the method without performing upstream data augmentation, using only a single modality for segmentation.

Metric		BraTS				$BraTS_s$			
Method		Dice	Sens	HD95	IoU	Dice	Sens	HD95	IoU
ratio1	w/o DA	0.6850	0.7339	3.7199	0.6496	0.6408	0.6451	4.8047	0.6201
	StarGAN v2	0.7057	0.7585	3.4459	0.6605	0.6463	0.6478	4.3161	0.6282
	AttGAN	0.6957	0.7444	3.5941	0.6590	0.6437	0.6441	4.6345	0.6249
	StarGAN v2 + CUT	0.7272	0.7664	3.0763	0.6968	0.6507	0.6631	4.0283	0.6316
	StarGAN v2 + QS	0.7460	0.7726	2.9412	0.7066	0.6693	0.6869	3.6507	0.6401
	U-GAT-IT	0.7302	0.7527	3.2106	0.6827	0.6536	0.6651	4.1140	0.6293
	AttentionGAN	0.7391	0.7690	3.1153	0.7004	0.6627	0.6772	3.8318	0.6390
	MSA-GAN	0.7415	0.7681	3.0367	0.6993	0.6641	0.6804	4.0637	0.6385
	SuperstarGAN	0.7509	0.7792	2.9210	0.7126	0.6685	0.6882	3.5913	0.6429
	W^2GAN	0.7476	0.7753	2.9578	0.7084	0.6705	0.6793	3.6152	0.6404
	AttCL-GAN (Ours)	**0.7551**	**0.7848**	**2.8937**	**0.7158**	**0.6797**	**0.6995**	**3.4525**	**0.6488**
ratio2	w/o DA	0.6850	0.7339	3.7199	0.6496	0.6408	0.6451	4.8047	0.6201
	StarGAN v2	0.7507	0.7661	3.1211	0.7120	0.6590	0.6923	4.1140	0.6395
	AttGAN	0.7446	0.7543	3.2463	0.7047	0.6485	0.6807	4.5585	0.6287
	StarGAN v2 + CUT	0.7673	0.7766	2.8975	0.7260	0.6634	0.7032	3.9740	0.6411
	StarGAN v2 + QS	0.7878	0.7888	2.7317	0.7505	0.6858	0.7150	3.5452	0.6506
	U-GAT-IT	0.7706	0.7792	3.0947	0.7301	0.6622	0.7048	4.0055	0.6399
	AttentionGAN	0.7817	0.7815	2.9799	0.7361	0.6773	0.7081	3.7396	0.6495
	MSA-GAN	0.7764	0.7684	3.0468	0.7448	0.6763	0.6994	3.7015	0.6458
	SuperstarGAN	0.7882	0.7892	2.8432	0.7504	0.6885	0.7185	3.5131	0.6514
	W^2GAN	0.7861	0.7856	2.6805	0.7492	0.6817	0.7093	3.4572	0.6492
	AttCL-GAN (Ours)	**0.7902**	**0.7949**	**2.6237**	**0.7523**	**0.6905**	**0.7217**	**3.2764**	**0.6528**
ratio3	w/o DA	0.6850	0.7339	3.7199	0.6496	0.6408	0.6451	4.8047	0.6201
	StarGAN v2	0.8048	0.8163	2.5925	0.7644	0.7360	0.7548	3.6658	0.6959
	AttGAN	0.7902	0.8072	2.6237	0.7523	0.7242	0.7464	3.9658	0.6907
	StarGAN v2 + CUT	0.8287	0.8297	2.2178	0.7822	0.7561	0.7741	3.4649	0.7154
	StarGAN v2 + QS	0.8293	0.8353	2.0969	0.7934	0.7701	0.7899	3.1611	0.7308
	U-GAT-IT	0.8171	0.8269	2.2439	0.7847	0.7629	0.7709	3.4525	0.7232
	AttentionGAN	0.8287	0.8322	2.1604	0.7919	0.7660	0.7860	3.2148	0.7240
	MSA-GAN	0.8264	0.8343	2.1796	0.7898	0.7547	0.7732	3.2691	0.7299
	SuperstarGAN	0.8330	0.8406	2.1062	0.7983	0.7763	0.7945	3.2064	0.7405
	W^2GAN	0.8295	0.8394	2.0031	0.7925	0.7725	0.7918	3.1437	0.7341
	AttCL-GAN (Ours)	**0.8391**	**0.8495**	**1.9880**	**0.8026**	**0.7882**	**0.8049**	**3.0178**	**0.7460**

translation. They also lack characterization of tissue and texture details. These are the two issues previously mentioned: poor lesion generation capability and insufficient tissue detail synthesis.

ii) StarGAN v2 + CUT and StarGAN v2 + QS achieve correct preservation of lesions in different translations by incorporating the idea of contrastive learning, but the lesion regions are sometimes imcomplete. Meanwhile, U-GAT-IT, AttentionGAN and MSA-GAN, as three cross-domain methods, have undergone three rounds of training and exhibit significant performance differences among all translations, with the boundary of lesions being blurred and showing poor robustness.

iii) Excluding AttCL-GAN, W^2GAN and SuperstarGAN show the best visual results among comparison models. They preserve structure and texture well across different modality translations, with no obvious semantic loss. Although still behind the top method in boundary clarity and detail completeness, as the latest baselines, their generated images are realistic and have certain advantages.

iv) The proposed AttCL-GAN firstly accomplishes stable generation of lesion regions, with the lesion positions and sizes being closest to the ground truth among seven methods. Secondly, our model ensures good generation of tissue details across different translations. Finally, the generated multi-modal images exhibit excellent discriminability, being easy to identify by the naked eye.

5.1.2. Downstream segmentation stage

After applying the aforementioned ten methods to complement modal missing images in the original dataset, we utilize these enhanced multi-modal datasets to evaluate performance improvements in segmentation models. Firstly, for every patient, we concatenate his images of each modality along the channel dimension, ensuring that every pixel position contains information from all modalities. Secondly, we adjust the existing segmentation model to adapt to the multi-modal input. This requires modifying the network's input channel number to ensure compatibility with the concatenated images. Finally, we train the segmentation model using the concatenated multi-modal images, allowing the model to learn representations for performing tasks across all modalities. And evaluate the segmentation performance of the model through the testing data.

Specifically, based on the six segmentation metrics in Table 3, we draw the following five observations:

i) The results of all multi-modal segmentation tasks are much higher than those achieved by training with single-modal data. This demonstrates the enhancing effect of multi-modal datasets on the performance of segmentation models and affirms the research significance of our work.

ii) Both StarGAN v2 + CUT and StarGAN v2 + QS outperform the two basic multi-domain translation baselines in all metrics, and StarGAN v2 + QS performs better than StarGAN v2 + CUT, which is consistent with the results of the upstream generation. On the one hand, it demonstrates that incorporating contrastive learning into multi-domain translation models can indeed enhance the quality of generated images and further improve the performance of downstream segmentation tasks. On the other hand, by calculating entropy to select anchors, the model can better focus on the information that

Table 4
The results of the ablation studies, where ACL represents the attention-guided contrastive learning module, AttAdaIN represents the attentional AdaIN-based generator, and SCD represents the style code diversity loss.

Metrics		BraTS				$BraTS_s$			
Methods		Dice	Sens	HD95	IoU	Dice	Sens	HD95	IoU
ratio1	StarGAN v2	0.7057	0.7585	3.4459	0.6605	0.6463	0.6478	4.3161	0.6282
	StarGAN v2 + ACL	0.7263	0.7634	3.3608	0.6894	0.6622	0.6668	3.8927	0.6349
	StarGAN v2 + AttAdaIN	0.7194	0.7592	3.4007	0.6773	0.6556	0.6591	4.0469	0.6310
	StarGAN v2 + SCD	0.7130	0.7589	3.4292	0.6755	0.6538	0.6569	4.1224	0.6299
	StarGAN v2 + ACL + AttAdaIN	0.7495	0.7760	3.0051	0.7117	0.6717	0.6890	3.5379	0.6423
	StarGAN v2 + ACL + SCD	0.7395	0.7699	3.2162	0.6979	0.6697	0.6737	3.6532	0.6417
	StarGAN v2 + AttAdaIN + SCD	0.7339	0.7656	3.3442	0.6958	0.6648	0.6712	3.8021	0.6387
	AttCL-GAN (Ours)	**0.7551**	**0.7848**	**2.8937**	**0.7158**	**0.6797**	**0.6995**	**3.4525**	**0.6488**

needs to be retained during translations, ensuring the generation of lesion areas, which is crucial for downstream segmentation.

iii) U-GAT-IT, AttentionGAN and MSA-GAN exhibit higher segmentation metrics compared to two basic multi-domain translation baselines, which is attributed to the use of attention mechanisms. This proves that by introducing attention mechanisms, the model can remedy the issue of insufficient tissue detail synthesis, generate higher-quality training data, and ultimately benefit downstream segmentation tasks. Additionally, the above three models showcase segmentation performance comparable to the two contrastive learning baselines, consistent with the quantitative results from the upstream generation experiments. As a result, repeated elaboration is omitted for conciseness.

iv) W^2GAN and SuperstarGAN achieve better segmentation results than the aforementioned seven methods, with performance second only to AttCL-GAN. W^2GAN enhances boundary and detail representation by generating new data. Its wavelet feature guidance emphasizes high-frequency features, while the importance weighting mechanism directs the model's attention to key regions, thereby improving the segmentation of complex lesions. SuperstarGAN maintains semantic consistency across modalities by using an independent classifier and data augmentation strategies, leading to more diverse and medically plausible images, which boost the overall performance of the segmentation model.

v) Our proposed AttCL-GAN has the state-of-the-art performance among the seven data augmentation methods. The reasons of superior performances of AttCL-GAN are as follows: Firstly, AttCL-GAN utilizes a Attention-Guided Contrastive Learning (ACL) module and a Attentional AdaIN-based (AttAdaIN) generator to address the issues of poor lesion generation capability and insufficient tissue detail synthesis respectively. Secondly, the experimental results comparing AttCL-GAN with two contrastive learning baselines demonstrate the superiority of the proposed ACL module, surpassing the other two contrastive learning approaches. Similarly, the experimental results comparing AttCL-GAN with three attention-based baselines validate that the proposed AttAdaIN generator performs better in leveraging attention mechanisms for medical images compared to the other two attention methods. Finally, AttCL-GAN additionally introduces a Style Code Diversity (SCD) loss to alleviate the high inter-modal similarity of results.

5.2. Ablation study

To show the effectiveness and necessity of the proposed three innovations in AttCL-GAN, ablation studies are further conducted, where several intermediate models that only use one or two advanced compoents are introduced and evaluated. Using StarGAN v2 + ACL as an example, it is derived from adding the attention-guided contrastive learning (ACL) module into StarGAN v2. The naming convention for other intermediate models follows the same pattern, so they are not elaborated further. Different from the main experiment, due to the downstream segmentation task being the ultimate application scenario of the proposed data augmentation method, in the ablation experiments, we only analyze the performance based on the quantitative results of the segmentation experiments.

In Table 4, all intermediate models outperform StarGAN v2 on four metrics, which demonstrates the effectiveness of the three proposed advancements in augmenting multi-modal datasets and ultimately improving segmentation performance. Specifically, we first compare the results of StarGAN v2 and StarGAN v2 + SCD, where StarGAN v2 + SCD outperforms StarGAN v2 in terms of all metrics. This is because style code is the numerical representation of style information in the latent space, containing style codes for corresponding and non-corresponding domains of the input. By maximizing the loss, we impose a new constraint on the generator G, enabling it to expand the latent distributions of different domains, ultimately achieve lower inter-modal similarity of results.

Then, it is observed that both StarGAN v2 + ACL and StarGAN v2 + AttAdaIN outperform StarGAN v2. This is because StarGAN v2 + ACL purposefully select patches of crucial regions using attention maps obtained from A. This enables the model to learn common information that needs to be retained and special information that should be ignored during modal translations, alleviating the issue of insufficient lesion generation. StarGAN v2 + AttAdaIN, on the other hand, enhances the generation over tissue structure and texture details by pixel-wise addition of the features extracted by the attention network A to the original features, and then inputting the combined into the AdaIN layer for normalization. These demonstrate the effectiveness of attention mechanism and contrastive learning in data augmentation for medical image.

Furthermore, we notice that StarGAN v2 + ACL + AttAdaIN is always better than StarGAN v2 + ACL and StarGAN v2 + AttAdaIN. This is because the two advancement, ACL and AttAdaIN, improve the generative capability of the base model on medical images by addressing different problems. This thus proves that it is reasonable to incorporate both ACL and AttAdaIN into the StarGAN v2 to enhance the generation performance and ultimately help the segmentation task to achieve more accurate results. Similar observations and conclusions are also obtained by comparing StarGAN v2 + ACL + SCD to StarGAN v2 + ACL and StarGAN v2 + SCD, comparing StarGAN v2 + AttAdaIN + SCD to StarGAN v2 + AttAdaIN and StarGAN v2 + SCD.

Finally, we find that AttCL-GAN constantly achieves much better performances than StarGAN v2 + ACL + AttAdaIN, StarGAN v2 + ACL + SCD and StarGAN v2 + AttAdaIN + SCD. This is because the three innovations target at addressing different issues, and as discovered by the first two findings, the three components can work together, complementing each other to enhance the model's generation capability, and ultimately improve segmentation performance. Therefore, the above observations demonstrate that the proposed three advancement are all effective and essential for data augmentation on medical multi-modal datasets using the backbone of StarGAN v2.

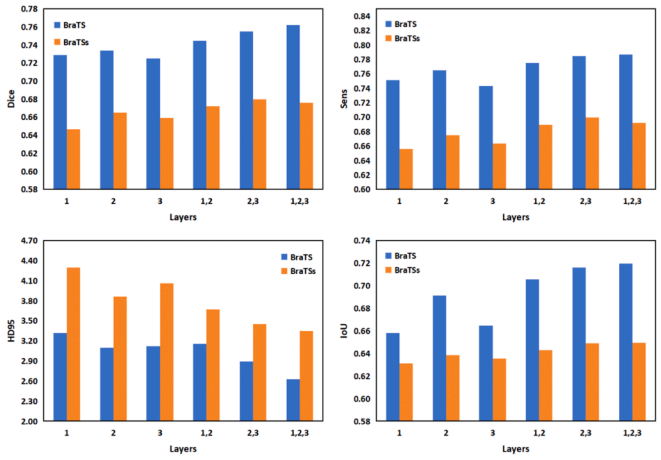


Fig. 4. Segmentation results in terms of four metrics on BraTS and BraTS$_s$ with varying hyper-parameter *Layers*. On the horizontal axis, *Layers* denotes the network layers on which the contrastive loss is calculated. Commas indicate simultaneous selection, enabling the choice of single, double, or triple layers.

5.3. Performances under different cases of modal data absence

In actual medical application scenarios, the modal absence in multi-modal datasets are diverse and complex. Thus, based on the selected BraTS dataset, we designed three cases with different degrees of modality missing. Detailed descriptions can be found in Table 1 of Section 4.1. In the experiments of Sections 5.1 and 5.2 above, we prioritized selecting ratio1 to evaluate the model performance. This is to obtain the performance of the model under the worst-case absence scenario, which is also the lower bound of performance among various scenarios. Consequently, in this section, we further supplemented the experiment of the nine baselines and AttCL-GAN and under cases of ratio2 and ratio3 to revalidate the superiority of the proposed model. Specific experimental results are shown in Table 3.

According to Table 3, we can draw the following conclusions. Firstly, from the perspective of different models, the result comparisons among single-modal segmentation without data augmentation, the nine baselines, and our proposed AttCL-GAN remain consistent with the four conclusions observed in Section 5.1.2. That is, both attention mechanism and contrastive learning approach are highly effective in improving the quality of multi-modal datasets and subsequently enhancing the performance of downstream segmentation tasks. Furthermore, under various modal absence scenarios, our proposed model consistently outperforms the others, achieving the best segmentation metrics. Secondly, from the perspective of different ratios, all seven models achieve the lowest segmentation results in the ratio1 scenario, better results in ratio2, and the best results in ratio3. This aligns with theoretical expectations: fewer missing multi-modal slices mean richer information about the four modalities, making it easier for the model to generate images with distinct modal characteristics while preserving tissue structure.

5.4. Effect of varying hyper-parameter *Layers*

The attention-guided contrastive learning module is the biggest innovation of the proposed AttCL-GAN. In our experiments, we use the hyper-parameter *Layers* to control the calculation of the contrastive loss at different layers. Our generator consists of a pair of four-layer and strcturally symmetric encoder and decoder. As different layers extract feature representations from distinct scales, calculating the contrastive loss at different layers results in varying qualities of generated images and improvements in segmentation performance. Consequently, experiments are conducted to investigate the effect of varying the hyper-parameter *Layers* on improving segmentation task performance.

Segmentation results in terms of four metrics on BraTS and BraTS$_s$ are shown in Fig. 4, through which we observed the following conclusions: (i) As for BraTS, the model achieved the best performance with *Layers* set to 1, 2, 3, and second best when the variable values were 2, 3. Compared to contrastive learning in a single layer and double layers, calculating losses on three layers can comprehensively utilize both low-level local features and high-level global structures, making it the ideal model design. However, due to high computational costs and increased training difficulties, we ultimately adopted a double-layer rather than a three-layer model design. (ii) The model performance on BraTS$_s$ is roughly consistent with that on BraTS. Overall, changes in this hyper-parameter has a smaller impact on performance, and the result of using double-layer is slightly better than that of three-layer. This is because the limited sample size weakens the constraint of the contrastive loss on G, failing to fully leverage the capabilities of contrastive learning.

5.5. Performances on small dataset approaching real-world scenarios

In real-world scenarios, abnormal images with lesions are often rare compared to normal images, especially in the field of oncology. Therefore, training models on small datasets to accomplish the data imputation is more applicable in medical practice. To validate the proposed model's ability to handle small-sample datasets in real clinical settings, we partitioned a subset from BraTS, denoted as BraTS$_s$. The detailed description of this partition is provided in Section 4.1 and Table 1.

In the main experiment, the generation performance on BraTS$_s$ is presented in Table 2, and the segmentation performance is shown in Table 3. Based on the results above, we observe the following two conclusions: (i) AttCL-GAN outperforms the other nine baselines in all translation processes in terms of all generation metrics and consistently achieves the best results in all segmentation metrics, demonstrating the effectiveness of our proposed model in real-world small-sample medical scenarios. (ii) As shown in Tables 2 and 3, each data augmentation method shows some performance differences between the original and the small-sample dataset. Specifically, in the generation task, except for the T1-to-Flair translation, the result differences are consistently the smallest for AttCL-GAN in all other translations. In the segmentation task, AttCL-GAN exhibits the smallest performance gaps in terms of Sens and HD95. As for the Dice and IoU metrics, the differences are the smallest in the ratio3 scenario. This proves the robustness of the proposed model under varying training sample sizes.

6. Discussion and future work

6.1. Social impact of proposed model

Although our study only focuses on medical segmentation tasks, in practice, most medical analysis models trained on multi-modal datasets require the datasets to be class-balanced, meaning an equal number of samples for each modality. Therefore, our designed multi-domain generative model can be applied to many clinical scenarios, where doctors or researchers may already have a multi-modal dataset with imbalanced class distribution or have collected several homogenous single-modal datasets with varying image quantities and hope to fully utilize these images for the subsequent analytical tasks. By using AttCL-GAN and the proposed modality completion strategy, we can perform image translations between any modalities and ultimately fill in the missing modal images when faced with the aforementioned practical challenges.

6.2. Limitation and future work

Based on the above discussion, although AttCL-GAN has addressed the modality missing in multi-modal datasets and improved the performance of downstream segmentation models, upon reviewing the entire research process, we have identified two problems with the proposed model.

Firstly, since our research focuses on achieving arbitrary translations between multiple (more than 2) modalities, we need to select homogenous multi-modal images to evaluate the model's performance. Such data can be sourced from specialized hospitals and medical institutions, which offer real clinical situations but are often unavailable due to privacy and ethical concerns. Alternatively, public platforms and large-scale medical imaging competitions provide relevant datasets like BraTS and IXI, containing four MRI modalities and focusing on brain structures and tumors. However, these datasets are relatively limited and do not fully simulate diverse clinical scenarios. Secondly, our attention network is implemented using the CBAM module. While this module enhances the network's perceptual capability by integrating channel and spatial attention information, it was initially proposed by Woo et al. [41] in 2018. Despite its significant achievements in past research, more flexible attention mechanisms have since emerged, and CBAM module may eventually be replaced in the future.

Therefore, further efforts to search for more multi-modal datasets beyond the brain, and validate the performance and generalization of the proposed model using them, could be the next step for improvement in this research. Additionally, introducing more innovative attention modules or replacing the GAN-based backbone with diffusion models, which can avoid mode collapse issues in GAN training, and making them suitable for medical images would also be an interesting future work.

7. Conclusion

In this work, we first identified three issues arising when applying StarGAN v2 to medical imaging: poor lesion generation capability, insufficient tissue detail synthesis and high inter-modal similarity of results. Subsequently, we proposed a novel attentional contrastive learning-based generative model, abbreviated as AttCL-GAN, to address these problems using three innovation: the attention-guided contrastive learning (ACL) module, the attentional AdaIN-based (AttAdaIN) generator and the style code diversity (SCD) loss. Furthermore, we introduced a modality completion strategy to intelligently perform data imputation for datasets with various modality missing situations using AttCL-GAN. Extensive experimental studies are conducted on a real-world multi-modal medical image datasets, and the results show that the proposed AttCL-GAN can significantly outperform the state-of-the-art GAN-based data augmentation methods in both medical image generation and segmentation tasks in terms of all metrics, and the three advancement are all effective and essential for AttCL-GAN to achieve the superior performance.

CRediT authorship contribution statement

Zhenghua Xu: Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Funding acquisition, Conceptualization; **Jiaqi Tang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis; **Dan Yao:** Writing – review & editing, Methodology; **Zhenzhen Wang:** Writing – review & editing, Visualization; **Thomas Lukasiewicz:** Writing – review & editing, Supervision, Funding acquisition.

Data availability

The data used are publicly available.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under the grant 62276089, by the Natural Science Foundation of Tianjin, China, under the grant 24JCJQJC00200, by the Natural Science Foundation of Hebei Province, China, under the grant F2024202064, by the Ministry of Human Resources and Social Security, China, under the grant RSTH-2023-135-1, and by the S&T Program of Hebei under the grant 24464401D. This work was also partially supported by the AXA Research Fund.

References

[1] J. Zhang, S. Zhang, X. Shen, T. Lukasiewicz, Z. Xu, Multi-ConDoS: multimodal contrastive domain sharing generative adversarial networks for self-supervised medical image segmentation, IEEE Trans. Med. Imaging 83 (2023) 102656.
[2] X. Fan, W. Zhou, X. Qian, W. Yan, Progressive adjacent-layer coordination symmetric cascade network for semantic segmentation of multimodal remote sensing images, Expert Syst. Appl. 238 (2024) 121999.
[3] J. Chen, J. Chen, Multimodal image feature fusion for improving medical ultrasound image segmentation, Biomed. Signal Process. Control 89 (2024) 105705.
[4] Z. Liu, Y. Cheng, T. Tan, T. Shinichi, MimicNet: mimicking manual delineation of human expert for brain tumor segmentation from multimodal MRIs, Appl. Soft Comput. 143 (2023) 110394.
[5] G. Chen, Q. Li, F. Shi, I. Rekik, Z. Pan, RFDCR: Automated brain lesion segmentation using cascaded random forests with dense conditional random fields, NeuroImage 211 (2020) 116620.
[6] G. Chen, J. Ru, Y. Zhou, I. Rekik, Z. Pan, X. Liu, Y. Lin, B. Lu, J. Shi, MTANS: multi-scale mean teacher combined adversarial network with shape-aware embedding for semi-supervised brain lesion segmentation, NeuroImage 244 (2021) 118568.
[7] X. Gu, Y. Zhang, W. Zeng, S. Zhong, H. Wang, D. Liang, Z. Li, Z. Hu, Cross-modality image translation: CT image synthesis of MR brain images using multi generative network with perceptual supervision, Comput. Methods Programs Biomed. 237 (2023) 107571.
[8] Y. Li, Q. Zheng, Y. Wang, Y. Zhou, Y. Zhang, Y. Song, W. Jiang, Fully convolutional transformer-based GAN for cross-modality CT to PET image synthesis, in: International Workshop on Computational Mathematics Modeling in Cancer Analysis, Springer, 2023, pp. 101–109.
[9] Z. Qin, Z. Liu, P. Zhu, W. Ling, Style transfer in conditional GANs for cross-modality synthesis of brain magnetic resonance images, Comput. Biol. Med. 148 (2022) 105928.
[10] Y. Choi, M. Choi, M. Kim, J.W. Ha, S. Kim, J. Choo, Stargan: unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[11] Y. Choi, Y. Uh, J. Yoo, J.W. Ha, Stargan v2: diverse image synthesis for multiple domains, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8188–8197.
[12] A. Tomczak, S. Ilic, G. Marquardt, T. Engel, F. Forster, N. Navab, S. Albarqouni, Multi-task multi-domain learning for digital staining and classification of leukocytes, IEEE Trans. Med. Imaging 40 (10) (2020) 2897–2910.
[13] Y. PENG, Z. MENG, L. YANG, Image-to-image translation for data augmentation on multimodal medical images, IEICE Trans. Inf. Syst. 106 (5) (2023) 686–696.
[14] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[15] M. Mirza, S. Osindero, Conditional generative adversarial nets, ttp://arxiv.org/abs/arXiv:1411.17841411.1784 (2014).
[16] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2223–2232.
[17] Z. Yi, H. Zhang, P. Tan, M. Gong, Dualgan: unsupervised dual learning for image-to-image translation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2849–2857.
[18] T. Kim, M. Cha, H. Kim, J.K. Lee, J. Kim, Learning to discover cross-domain relations with generative adversarial networks, in: International Conference on Machine Learning, PMLR, 2017, pp. 1857–1865.
[19] M.Y. Liu, T. Breuel, J. Kautz, Unsupervised image-to-image translation networks, Adv. Neural Inf. Process. Syst. 30 (2017).
[20] Z. He, W. Zuo, M. Kan, S. Shan, X. Chen, Attgan: facial attribute editing by only changing what you want, IEEE Trans. Image Process. 28 (11) (2019) 5464–5478.
[21] Q. Yang, Y. Pu, Z. Zhao, D. Xu, S. Li, W2GAN: Importance weight and wavelet feature guided image-to-image translation under limited data, Comput. Graph. 116 (2023) 115–127.
[22] K. Ko, T. Yeom, M. Lee, Superstargan: generative adversarial networks for image-to-image translation in large-scale domains, Neural Netw. 162 (2023) 330–339.
[23] B. Han, M. Hu, The facial expression data enhancement method induced by improved starGAN V2, Symmetry 15 (4) (2023) 956.
[24] Y.A. Li, A. Zare, N. Mesgarani, Starganv2-vc: a diverse, unsupervised, non-parallel framework for natural-sounding voice conversion, ttp://arxiv.org/abs/arXiv:2107.103942107.10394 (2021).

[25] K. He, H. Fan, Y. Wu, S. Xie, R. Girshick, Momentum contrast for unsupervised visual representation learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9729–9738.
[26] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: International Conference on Machine Learning, PMLR, 2020, pp. 1597–1607.
[27] W. Li, P. Zhang, L. Zhang, Q. Huang, X. He, S. Lyu, J. Gao, Object-driven text-to-image synthesis via adversarial training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 12174–12182.
[28] S. Zhang, J. Zhang, B. Tian, T. Lukasiewicz, Z. Xu, Multi-modal contrastive mutual learning and pseudo-label re-learning for semi-supervised medical image segmentation, Med. Image Anal. 83 (2023) 102656.
[29] K.S. Lee, N.T. Tran, N.M. Cheung, Infomax-gan: improved adversarial image generation via information maximization and contrastive learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 3942–3952.
[30] Z. Zhao, Z. Zhang, T. Chen, S. Singh, H. Zhang, Image augmentations for gan training, ttp://arxiv.org/abs/arXiv:2006.025952006.02595 (2020).
[31] Y. Deng, J. Yang, D. Chen, F. Wen, X. Tong, Disentangled and controllable face image generation via 3d imitative-contrastive learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5154–5163.
[32] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, H. Wang, Geometry-contrastive gan for facial expression transfer, ttp://arxiv.org/abs/arXiv:1802.018221802.01822 (2018).
[33] M. Kang, J. Park, Contragan: contrastive learning for conditional image generation, Adv. Neural Inf. Process. Syst. 33 (2020) 21357–21369.
[34] T. Park, A.A. Efros, R. Zhang, J.Y. Zhu, Contrastive learning for unpaired image-to-image translation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16, Springer, 2020, pp. 319–345.
[35] X. Hu, X. Zhou, Q. Huang, Z. Shi, L. Sun, Q. Li, Qs-attn: query-selected attention for contrastive learning in i2i translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 18291–18300.
[36] R.A. Rensink, The dynamic representation of scenes, Vis. Cogn. 7 (1–3) (2000) 17–42.
[37] J. Kim, M. Kim, H. Kang, K. Lee, U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation, ttp://arxiv.org/abs/arXiv:1907.108301907.10830 (2019).
[38] H. Tang, H. Liu, D. Xu, P.H.S. Torr, N. Sebe, Attentiongan: unpaired image-to-image translation using attention-guided generative adversarial networks, IEEE Trans. Neural Netw. Learn. Syst. 34 (4) (2021) 1972–1987.
[39] F. Wang, Q. Zhang, Q. Zhao, M. Wang, F. Sun, Unsupervised image-to-image translation with multiscale attention generative adversarial network, Appl. Intell. 54 (8) (2024) 6558–6578.
[40] S.H. Gao, M.M. Cheng, K. Zhao, X.Y. Zhang, M.H. Yang, P. Torr, Res2net: a new multi-scale backbone architecture, IEEE Trans. Pattern Anal. Mach. Intell. 43 (2) (2019) 652–662.
[41] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, Cbam: convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.
[42] X. Huang, M.Y. Liu, S. Belongie, J. Kautz, Multimodal unsupervised image-to-image translation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 172–189.
[43] R. Mechrez, I. Talmi, F. Shama, L. Zelnik-Manor, Maintaining natural image statistics with the contextual loss, in: Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14, Springer, 2019, pp. 427–443.
[44] X. Huang, S. Belongie, Arbitrary style transfer in real-time with adaptive instance normalization, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1501–1510.
[45] S. Liu, T. Lin, D. He, F. Li, M. Wang, X. Li, Z. Sun, Q. Li, E. Ding, Adaattn: revisit attention mechanism in arbitrary neural style transfer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 6649–6658.
[46] T. Park, M.Y. Liu, T.C. Wang, J.Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.
[47] Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 21–29.
[48] Q. Mao, H.Y. Lee, H.Y. Tseng, S. Ma, M.H. Yang, Mode seeking generative adversarial networks for diverse image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1429–1437.
[49] U. Baid, S. Ghodasara, S. Mohan, M. Bilello, E. Calabrese, E. Colak, K. Farahani, J. Kalpathy-Cramer, F.C. Kitamura, S. Pati, et al., The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, ttp://arxiv.org/abs/arXiv:2107.023142107.02314 (2021).