

Article

Unsupervised Head PD-to-T2 MR Image Translation via Multi-Scale Feature Regularization

Xu Chen ^{1,2,3,*} , Yuntian Bai ¹ and Yifeng Hong ¹ 

¹ College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China; 25013083001@stu.hqu.edu.cn (Y.B.); 24013083006@stu.hqu.edu.cn (Y.H.)

² Key Laboratory of Computer Vision and Machine Learning (Huaqiao University), Fujian Province University, Xiamen 361021, China

³ Xiamen Key Laboratory of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen 361021, China

* Correspondence: chenxu@hqu.edu.cn

Abstract

Unsupervised medical image translation remains challenging because model development often relies on unpaired training, whereas reliable evaluation requires well-matched reference images. PD-weighted and T2-weighted brain MR images provide a useful testbed for this problem because they are closely matched anatomically while still exhibiting distinct contrast characteristics. Existing methods often align only high-level features, overlooking low-level texture details that are important for structural fidelity. In this work, we propose the Multi-Scale Feature Regularization and Patch Mixup (MSFRPM) framework based on an encoder–decoder architecture. It aligns cross-domain features across multiple scales to preserve local details and employs a patch-based mixup strategy to augment training data. The framework was evaluated using an unsupervised learning protocol with strict data partitioning. Experimental results demonstrate that MSFRPM achieves strong performance relative to eight state-of-the-art methods. Our approach achieved improvements in MAE (6.26 ± 0.86), PSNR (23.53 ± 0.92), SSIM (0.83 ± 0.03), and GMSD (0.100 ± 0.010). Qualitative assessments confirmed improved structural fidelity, and t-SNE visualization validated enhanced cross-domain feature alignment. Overall, MSFRPM provides a useful approach for unsupervised PD-to-T2 image translation under the current experimental setting.

Keywords: cross-domain image translation; multi-scale feature regularization; patch mixup



Academic Editor: Giorgio Maria Di Nunzio

Received: 27 March 2026

Revised: 29 April 2026

Accepted: 9 May 2026

Published: 12 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

In brain MRI, multi-sequence imaging provides complementary contrast information for clinical reading and technical analysis. Proton Density (PD)-weighted and T2-weighted images serve distinct purposes, and PD/T2 pairs are particularly well aligned because they can be derived from the same spin-echo acquisition. This characteristic makes PD-to-T2 translation a suitable testbed for studying contrast transformation methods with minimal anatomical mismatch. At the same time, image translation between related MR contrasts remains meaningful because preserving low-level texture and structural fidelity in the synthesized contrast is still challenging.

Unsupervised image translation remains an important research direction because many cross-domain translation frameworks are designed for unpaired training and aim to learn robust domain-invariant anatomical representations without relying on direct pixel-wise supervision. Common methods train models to extract shared anatomical

features from different sequences using adversarial learning [1–4]. These methods align feature spaces across different domains. The model encodes source domain images into embeddings, which the target domain decoder then uses to synthesize images. For instance, UNIT [2] maps images from different domains into a shared latent space. DRIT++ [4] embeds images in a domain-invariant content space and a domain-specific attribute space. However, these methods often focus on aligning only the highest-level latent spaces, which may overlook low-level local texture details because those aligned features have relatively low spatial resolution.

In this study, we propose the Multi-Scale Feature Regularization and Patch Mixup (MSFRPM) method to tackle the unsupervised cross-domain medical image translation challenge, specifically targeting head PD-to-T2 MR image translation. Our model comprises a shared encoder for building domain-shared latent spaces, two domain-specific decoders to synthesize images for the respective domains, and two image discriminators. Unlike existing approaches, MSFRPM performs feature space alignment across multiple scales. Specifically, the intermediate layers of the encoder and decoders are categorized into low-level and high-level groups based on resolution. The high-level intermediate layers of the decoders are shared across domains and connected to the corresponding encoder layers via shortcut connections to mitigate local detail leakage. Instead of solely aligning the highest-level latent space, we propose aligning multiple feature spaces built by the encoder's high-level group across various scales. The low-level intermediate layers of the decoders are shortcut-connected to the discriminators' corresponding layers to facilitate adversarial learning across multiple scales concurrently. Additionally, we introduce a patch mixup approach that blends real images with synthetic ones to generate novel samples incorporating texture details from different domains. By regularizing the multi-scale features of mixup samples, the encoder effectively extracts domain-invariant representations. We evaluated MSFRPM on the IXI dataset for unsupervised PD-to-T2 image translation.

To summarize, this work makes the following key contributions:

- (1) We propose a multi-scale feature regularization method for unsupervised cross-domain medical image translation. Unlike conventional approaches that focus solely on aligning high-level representations, our method aligns feature spaces across multiple scales simultaneously. This strategy preserves more low-level local details during feature embedding, thereby enhancing the structural fidelity of synthesized images.
- (2) We introduce a patch-based mixup strategy to blend real images with synthetic ones. Regularizing the multi-scale features of mixup samples effectively improves our model's ability to extract domain-invariant features, ensuring robustness against texture variations.
- (3) Experimental results on unsupervised head PD-to-T2 translation demonstrate the efficacy of our model.

2. Related Work

2.1. Cross-Domain Image Translation

Cross-domain image translation involves transforming images from one domain to another using computational techniques such as deep learning. Generative Adversarial Networks (GANs) have significantly advanced this field, enabling more flexible, robust, and high-fidelity translations across different imaging modalities. Current methods in cross-domain image translation are primarily divided into two categories: image-level and feature-level. Image-level methods focus on learning the direct mappings from the pixels or voxels of a source image to those of a target image [5–9]. For example, Hiasa et al. [7] extended the CycleGAN framework [5] to enhance boundary accuracy in MR-to-CT synthesis by introducing a gradient consistency loss. Armanious et al. [8] developed Cycle-MedGAN,

which utilizes a cycle-consistency loss to facilitate training for cross-domain image translation without paired images. Additionally, Gu et al. [6] introduced a hybrid approach combining cycle consistency and style transfer to achieve structurally aligned MRI-to-CT transformations. Feature-level methods typically involve embedding domain-agnostic features from images of different domains using domain adaptation techniques [1–4,10]. For example, Liu et al. [2] proposed an unsupervised image-to-image translation framework (UNIT) based on the shared-latent space assumption. Lee et al. [4] introduced a method for embedding images onto a domain-invariant content space and a domain-specific attribute space, facilitating cross-domain translation by combining content features extracted from an input image with attribute features sampled from the attribute space. Additionally, Han et al. [10] proposed a dual learning framework to enhance inference capabilities in cross-domain settings.

Most existing feature-level cross-domain image translation methods predominantly focus on aligning high-level representations. However, these approaches often overlooks crucial low-level tissue details. In our study, we introduce a multi-scale feature regularization technique designed to effectively improve the quality of synthetic images in unsupervised cross-domain medical image translation.

2.2. Mixup Technique

The concept of mixup was first introduced by Zhang et al. [11] as a data augmentation technique that constructs virtual training examples by forming convex combinations of pairs of examples and their labels. This approach has been shown to enhance the generalization of neural networks by promoting simple linear behaviors between training samples. Subsequent studies have further developed the utility and theoretical foundations of mixup [12–17]. For instance, Yun et al. [12] introduced CutMix, which mixes training images by replacing patches of one image with those from another, enhancing performance in tasks like object detection. Guo et al. [13] discussed mixup's role in enhancing robustness and improving uncertainty estimation, noting its effectiveness as a regularizer against corrupt labels. Verma et al. [14] extended mixup to sequence models with Manifold Mixup, which mixes in both the input space and hidden layers, showing benefits in robustness and model interpretability for natural language and time series applications. Faramarzi et al. [15] proposed PatchUP, a feature-space block-level regularization technique for CNNs, switching contiguous blocks of feature maps between samples.

Overall, the mixup technique and its variants mark a significant advancement in machine learning, providing versatile tools for enhancing model training through data augmentation. Currently, MixUp methods are predominantly utilized in classification or segmentation tasks, with limited application in image generation. In this work, we introduce a patch-level mixup strategy, demonstrating that mixup can also effectively benefit cross-domain image translation tasks.

3. Materials and Methods

3.1. Dataset

The experimental data were sourced from the IXI dataset (<https://brain-development.org/ixi-dataset/>, accessed on 8 May 2026), a publicly available repository comprising multi-modal MR images of healthy subjects. This study focused on PD-weighted and T2-weighted sequences to evaluate the proposed cross-domain translation framework. The IXI dataset was originally collected under ethical approval from the local research ethics committee, and all subject data were anonymized prior to public release.

For evaluation of unsupervised translation performance, we implemented a data partitioning strategy that prevents information leakage between domains. Specifically,

we identified subjects who possessed both PD and T2 sequences to construct the paired validation and testing sets. From this pool, 100 subject-matched PD-T2 pairs were randomly selected for the testing set, and an additional 50 subject-matched pairs were allocated for the validation set. These paired samples were used only for quantitative evaluation and qualitative visual comparison.

For training, we ensured no subject overlap between the source (PD) and target (T2) domains. We extracted the remaining subjects from the repository, selecting 215 PD images and 214 T2 images from distinct participants. This disjoint sampling strategy allows the model to learn domain-invariant features without using paired structural information during training. Although paired PD/T2 data are available in this imaging setting, they were not used for model training in this study.

3.2. Preprocessing

To ensure data consistency and facilitate accurate evaluation, a comprehensive preprocessing pipeline was applied to all datasets.

For the paired validation and testing sets, each PD image was rigidly aligned with its corresponding T2 image using the NiftyReg software [18] v1.3.9 to correct for inter-scan motion and ensure pixel-wise correspondence for quantitative metrics. For the unpaired training sets, although subject-matched pairs were unavailable, all images were independently registered to a common standard space to ensure consistent anatomical orientation and spatial alignment across the dataset.

Subsequently, all images were corrected for bias field inhomogeneity using N4 to address low-frequency intensity inconsistencies inherent in MR acquisitions. To standardize the intensity distribution, we excluded the top 1% of intensity values in each volume to remove outliers and linearly scaled the voxel intensities to a range of $[-1, 1]$. Finally, all images were resized to a uniform resolution of 256×256 pixels.

3.3. Method Overview

In the unsupervised cross-domain image translation task, we have two image sets, $\mathcal{X}_S$ and $\mathcal{X}_T$, consisting of unpaired training samples from the source and target domains, respectively. In this work, we propose a Multi-Scale Feature Regularization and Patch Mixup (MSFRPM) method, which aims to learn a mapping from source domain images to target domain images using these unpaired image sets $\mathcal{X}_S$ (PD) and $\mathcal{X}_T$ (T2).

In image processing, the shortcut connection technique introduced in UNet [19] has proven highly effective for dense prediction tasks such as image generation and segmentation. It facilitates the direct transfer of features from earlier layers to later ones, preserving fine-grained details and enhancing the model's ability to recover spatial information. However, this technique is rarely employed in feature-level domain adaptation methods, as these methods typically apply feature regularization only to the highest-level representations, while shortcut connections can bypass the scale at which regularization is applied.

In this work, we introduce a multi-scale feature regularization approach. As shown in Figure 1, our model incorporates two generators, each with an encoder–decoder structure, to learn mappings between source-to-target and target-to-source. Traditionally, images during the encoding process yield multi-scale features at progressively decreasing resolutions. Typically, features at higher resolutions are regarded as low-level, while those at lower resolutions are considered high-level. We categorize the intermediate layers of both encoders and decoders into high-level and low-level groups, with the former handling domain-invariant features and the latter handling domain-specific features. Shortcut connections link the high-level intermediate layers of the encoder and decoder, similar to the UNet architecture. This setup allows for feature-level domain adaptation across multi-

scale feature spaces, rather than focusing exclusively on the highest-level representations. Additionally, the intermediate layers of the decoder's low-level group are connected to the discriminator's intermediate layers, enabling multi-scale regularization. This design preserves fine-grained details during regularization and enhances gradient flow across different scales. Further details of this approach are discussed in Section 3.4.

Furthermore, we propose a patch-based mixup approach, blending real and synthetic images to create novel samples that combine texture details from different domains. By regularizing the features of these mixed samples, we enhance the encoder's ability to extract domain-invariant representations. Details of this approach are provided in Section 3.5. Network structures and training details are elaborated in Section 3.6.

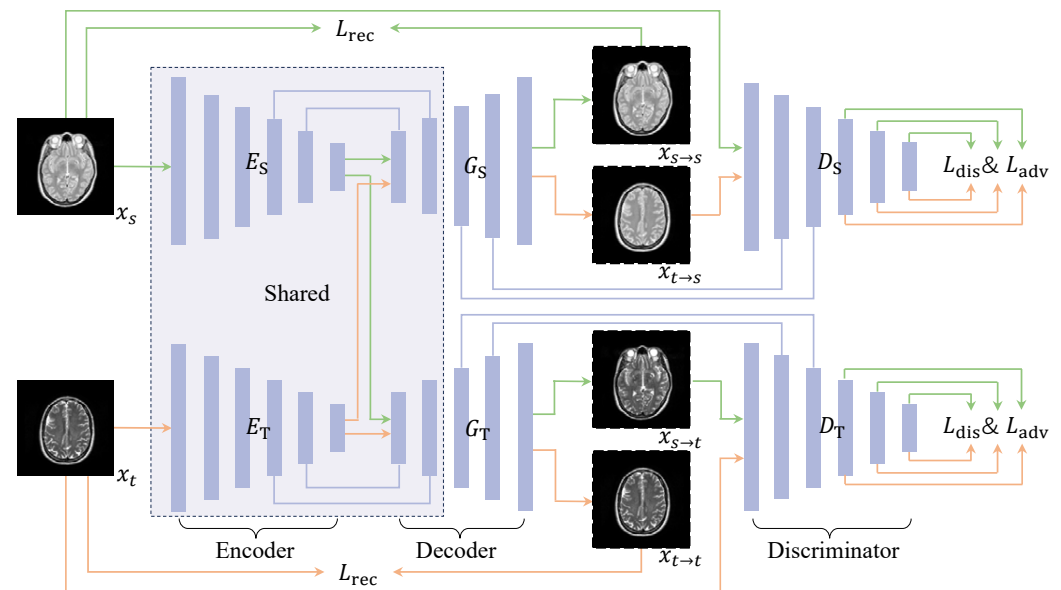


Figure 1. Illustration of framework. The intermediate layers of the encoder and decoder are split into the high-level group and low-level group. The former is shared across domains and shortcut-connected. The latter is domain-specific and be shortcut-connected with discriminator's corresponding layers. The green and orange arrows represent the flow paths of data from the source domain and the target domain, respectively.

3.4. Multi-Scale Feature Regularization

Let E_S and E_T represent the encoders for the source and target domains, respectively; G_S and G_T denote the decoders for these domains; and D_S and D_T are the discriminators for the source and target domains, respectively. Traditionally, the encoder/discriminator consists of several encoding convolutional blocks that progressively downsample the features, while the decoder is made up of decoding convolutional blocks that gradually upsample the features. In this work, we categorize the intermediate encoding and decoding blocks of the encoder/decoder/discriminator into low-level and high-level groups. We denote $P^H(\cdot)$ and $P^L(\cdot)$ as the outputs produced by the intermediate blocks of net P in the high-level and low-level groups, respectively, where $P \in \{E_S, E_T, G_S, G_T, D_S, D_T\}$.

The encoder and decoder are responsible for embedding the image into low-dimensional features and reconstructing the image from the features, respectively. Firstly, the encoder and decoder of the same domain server as an autoencoder, which are trained using the following self reconstruction function:

$$L_{\text{rec}} = \left\| G_S(E_S^H(x_s)) - x_s \right\|_1 + \left\| G_T(E_T^H(x_t)) - x_t \right\|_1, \quad (1)$$

where $x_s \in \mathcal{X}_S$ and $x_t \in \mathcal{X}_T$ represent the training samples from the source and target domains, respectively. Traditionally, in encoder–decoder network architectures, only the output of the final layer of the encoder is fed into the decoder. In UNet networks, however, the outputs of each encoding block of the encoder are used as inputs to the decoder. In our network, by contrast, only the outputs from high-level encoding blocks are utilized as multi-scale inputs to the decoder.

To enable source-to-target image translation, the features extracted from a source image x_s need to be utilized for generating a corresponding target image $x_{s \rightarrow t}$ through the decoder G_T of the target domain, represented as $x_{s \rightarrow t} = G_T(E_S^H(x_s))$. Conversely, for target-to-source translation, starting with the features extracted from a target image x_t , we can reconstruct a source image $x_{t \rightarrow s}$ using the decoder G_S of the source domain, as $x_{t \rightarrow s} = G_S(E_T^H(x_t))$. These reciprocal image constructions are termed as cross-reconstruction. To enhance the quality of cross-reconstructed images, we employ adversarial learning. The corresponding discrimination loss and adversarial loss are:

$$L_{\text{dis}} = (D_S^H(x_s) - 1)^2 + D_S^H(G_S^L(E_T^H(x_t)))^2 + (D_T^H(x_t) - 1)^2 + D_T^H(G_T^L(E_S^H(x_s)))^2, \quad (2)$$

$$L_{\text{adv}} = (D_S^H(G_S^L(E_T^H(x_t))) - 1)^2 + (D_T^H(G_T^L(E_S^H(x_s))) - 1)^2. \quad (3)$$

Unlike traditional discriminators that take only an image as input, our network integrates shortcut connections between the low-level intermediate layers of both the decoder and the discriminator. This allows adversarial learning to be applied not only to the images but also to the domain-specific low-level features generated by these intermediate layers. Specifically, for real images, the low-level features from the self-reconstruction process are fed into the discriminator, while for synthetic images, the low-level features from the cross-reconstruction process are also introduced. Moreover, whereas traditional discriminators produce output solely at the final layer, our discriminator generates multiple outputs from the high-level intermediate layers. Consequently, the discrimination loss function (Equation (2)) and adversarial loss function (Equation (3)) are applied to each output of the discriminator. This approach enables gradient flow at multiple scales, which effectively mitigates the instability issues often associated with adversarial learning [20].

Naturally, the cross-reconstructed image should exhibit anatomical features that are consistent with those of the real image. However, images from different domains inherently have distinct appearances. Therefore, it is not appropriate to demand that low-level features remain consistent across different domains. As illustrated in Figure 2, we propose a multi-scale cross-reconstruction feature consistency loss that enforces the consistency constraint specifically at high-level feature spaces. This is formulated as following:

$$L_{\text{fc}} = \|E_S^H(x_s) - E_T^H(G_T(E_S^H(x_s)))\|_2^2 + \|E_T^H(x_t) - E_S^H(G_S(E_T^H(x_t)))\|_2^2, \quad (4)$$

3.5. Patch-Based Mixup Strategy

In our framework, the shared encoder is designed to map images from different domains into a common latent space. To achieve this, we introduce a patch-based mixup strategy to enhance the encoder's ability to extract consistent features despite varying appearances. First, a binary mask M , sized identically to the training image, is generated. Following the implementation principle of PatchUP [15], we first sample candidate block centers randomly on the image grid and then expand each sampled center into a contiguous

square region. More specifically, let η denote the block edge length. For each spatial location, we perform Bernoulli sampling with adjusted probability

$$\gamma_{\text{adj}} = \frac{\gamma HW}{\eta^2(W - \eta + 1)(H - \eta + 1)}, \quad (5)$$

where γ_{adj} denotes the adjusted Bernoulli sampling probability, and H and W denote the height and width of the image grid used to generate the mask, respectively. The sampled map is then expanded by max-pooling with kernel size η , producing contiguous square holes. Finally, the binary mask is obtained as $M = 1 - \text{holes}$, where $M = 0$ indicates the replaced region and $M = 1$ indicates the preserved region. Therefore, γ controls the expected masking ratio, while η controls the block size. The same spatial mask pattern is shared across channels, and a new mask is randomly generated at each training iteration.

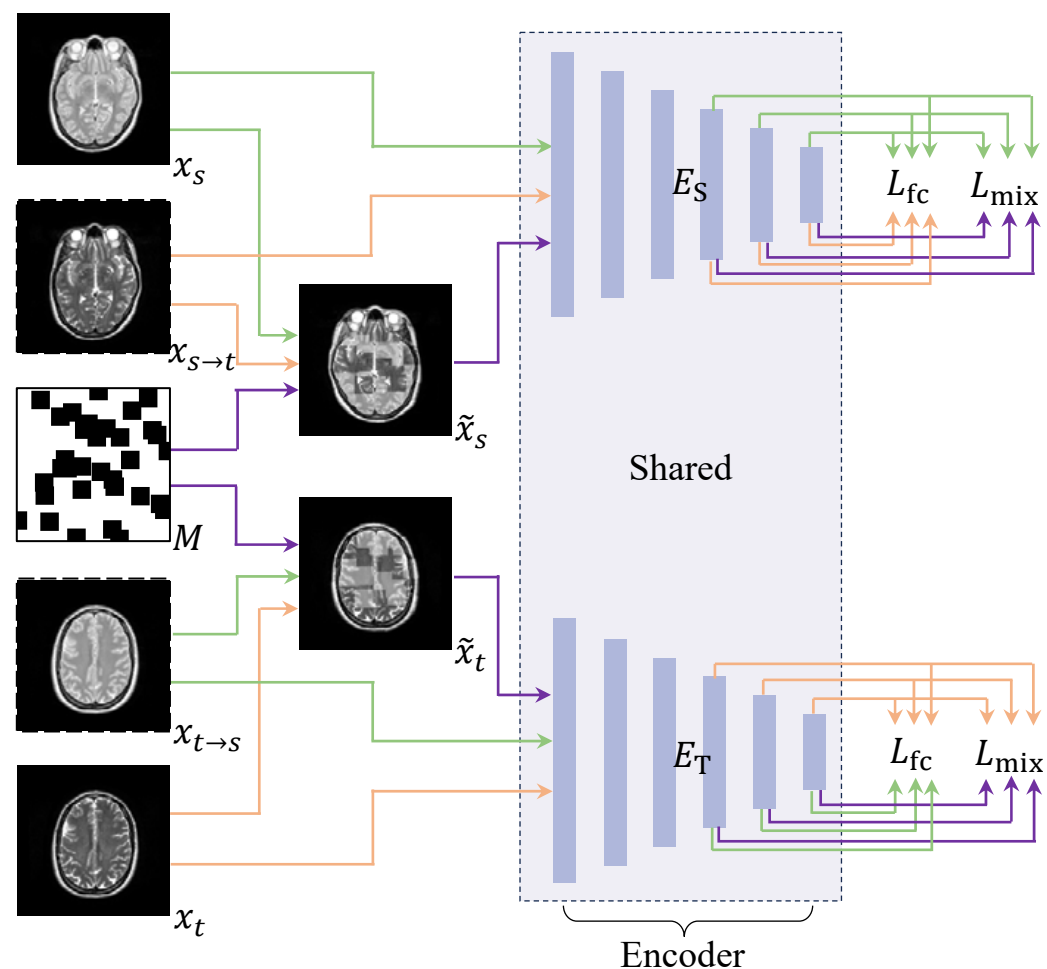


Figure 2. Illustration of multi-scale feature regularization approach and patch mixup strategy. The green and orange arrows represent the flow paths of data from the source domain and the target domain, respectively.

However, unlike PatchUP [15], which uses the generated mask to mix feature maps, we implement patch-based mixup at the image level. Specifically, a real image x_s and its corresponding cross-reconstructed image $x_{s \rightarrow t}$ form a pair of images from different domains, as well as x_t and $x_{t \rightarrow s}$. In this work, we mix a real image with its corresponding cross-reconstructed image to generate a new sample. For instance, $\tilde{x}_s = M \odot x_s + (1 - M) \odot x_{s \rightarrow t}$ and $\tilde{x}_t = M \odot x_t + (1 - M) \odot x_{t \rightarrow s}$, where $\odot$ denotes element-wise multiplication. Since the images used for mixing share the same anatomical structure, the anatomical features

of the mixed images should remain consistent. This regularization is enforced using the following mixup consistency loss function:

$$L_{\text{mix}} = \left\| E_S^H(x_s) - E_S^H(\tilde{x}_s) \right\|_2^2 + \left\| E_T^H(x_t) - E_T^H(\tilde{x}_t) \right\|_2^2, \quad (6)$$

By utilizing mixed images as augmented training data, the model becomes more resilient to cross-domain variations and distortions, as it is trained on samples combining tissues from different domains. Additionally, our patch-based mixup strategy serves as a regularizer, mitigating overfitting by introducing greater variability into the training data.

3.6. Other Details

3.6.1. Network Architecture

Our encoder consists of an initial 3×3 convolutional layer, a residual block [21], and five encoding convolutional blocks. Each encoding block contains a 3×3 convolutional layer with a stride of 2 for downsampling, followed by a residual block. The decoder consists of five decoding convolutional blocks and an additional 3×3 convolutional layer to generate the synthesized image. Each decoding block starts with a transposed convolutional layer with a stride of 2 to upsample the feature map, followed by a residual block. The discriminator has a structure similar to the encoder, but with each encoding block in the high-level group augmented by a 3×3 convolutional layer to produce intermediate outputs for multi-scale adversarial learning.

As illustrated in Figure 1, the encoding and decoding convolutional blocks are categorized into high-level and low-level groups. In this framework, the encoders and the intermediate layers of the decoders in the high-level group share weights across domains. Shortcut connections, as proposed in UNet [19], are used to link the high-level blocks of the encoder and decoder, as well as to connect the low-level blocks of the decoder and discriminator, as depicted in Figure 1. The grouping into high-level and low-level groups is detailed in the Results section. Finally, Group Normalization [22] is applied after each convolutional and transposed convolutional layer, which has been shown to effectively stabilize and accelerate the training process. For activation functions, Swish [23] is used for the encoder and decoder, while Leaky ReLU with a slope of 0.2 is used for the discriminator.

3.6.2. Training Details

Our model is trained adversarially. Specifically, during each training iteration, the discriminators D_S and D_T are first updated using Equation (2) with the encoders and decoders held constant. Following this, the discriminators are kept fixed while the encoders and decoders are trained using the combined loss function:

$$L_{\text{rec}} + \lambda_1 L_{\text{adv}} + \lambda_2 L_{\text{fc}} + \lambda_3 L_{\text{mix}}, \quad (7)$$

We trained the model for 50 epochs on an NVIDIA GeForce RTX 4090 GPU using the AdamW optimizer [24]. The learning rate was set to 0.0001, and the batch size was 1. The weights for the loss function were configured as $\lambda_1 = 0.01$, $\lambda_2 = 0.1$, and $\lambda_3 = 1$, based on evaluation results from the validation set.

4. Results

4.1. Quantitative and Qualitative Results

During the evaluation phase, we compared our MSFRPM with eight unsupervised cross-domain image translation methods: CycleGAN [5], Swin-UNIT [25], ContourDiff [26], DCLGAN [10], DRIT++ [4], NICE-GAN [27], MSA-GAN [28], and SynDiff [29]. We re-trained these comparison methods using their officially published source code under the

same preprocessing pipeline, the same train/validation/test partition strategy, and the same validation-based model selection criterion as used for MSFRPM. For evaluation, we used widely recognized metrics: Mean Absolute Error (MAE), which measures pixel-wise absolute difference; Peak Signal-to-Noise Ratio (PSNR), which reflects reconstruction fidelity based on intensity error; Structural Similarity Index Metric (SSIM), which evaluates perceived structural similarity; and Gradient Magnitude Similarity Deviation (GMSD), which provides a gradient-based assessment of structural distortion. Similar to MSFRPM, all comparison methods were periodically evaluated on the validation set during training, and the checkpoint with the highest validation PSNR was selected for testing. We chose PSNR as the common model-selection criterion for all methods because it is a standard measure of reconstruction fidelity in image translation and allowed us to apply a consistent validation protocol across methods.

The quantitative performance of all methods is detailed in Table 1. Our MSFRPM consistently achieves the best results across all metrics. Specifically, the performance improvements over the second-best method are 0.77 in MAE, 1.67 in PSNR, 0.03 in SSIM, and 0.012 in GMSD. In terms of GMSD, Swin-UNIT and DCLGAN are the strongest competing methods, followed closely by SynDiff, whereas CycleGAN, MSA-GAN, NICE-GAN, DRIT++, and ContourDiff show noticeably worse structural fidelity. For statistical testing, one sample corresponded to one subject-matched PD–T2 pair in the test set, and the image quality metrics were computed separately for each subject pair. We then used paired *t*-tests to compare the per-subject metric values of each competing method against those of MSFRPM. Additionally, the symbols * and ** in the tables indicate that the observed improvements are statistically significant at the 5% and 1% levels, respectively, reflecting a low probability that these performance gains are due to chance.

The representative visual results are presented in Figure 3, where each subfigure displays: (a) the input PD-weighted MR image, (b–j) the translation results from nine comparison methods, (k) our MSFRPM result, and (l) the ground-truth T2-weighted image. Additionally, the images in the even rows show heatmaps that reflect the difference between the synthetic images produced by each method and the ground truth. Warmer colors (red) indicate larger differences, while cooler colors (blue) represent smaller differences. Our MSFRPM shows more favorable qualitative results than the comparison methods in many regions, as evidenced by the cooler colors in most areas of its error heatmaps, suggesting that the translated images are closer to the target images. These qualitative results align with the quantitative findings.

Table 1. Quantitative results of head PD-to-T2 MR image translation. Best results are marked in bold. * and ** denote *p*-value < 0.05 and *p*-value < 0.01, respectively.

Methods	MAE ($\times 100$) ↓	PSNR ↑	SSIM ↑	GMSD ↓
CycleGAN	8.09 ± 0.61 **	21.28 ± 0.47 **	0.78 ± 0.03 **	0.137 ± 0.003 **
Swin-UNIT	7.79 ± 0.75 **	21.61 ± 0.70 **	0.81 ± 0.02 *	0.112 ± 0.008 **
ContourDiff	15.37 ± 1.34 **	16.93 ± 0.60 **	0.42 ± 0.04 **	0.161 ± 0.008 **
DCLGAN	7.03 ± 0.32 **	21.86 ± 0.30 **	0.80 ± 0.02 **	0.113 ± 0.005 **
DRIT++	9.97 ± 0.73 **	19.86 ± 0.54 **	0.73 ± 0.05 **	0.158 ± 0.007 **
NICE-GAN	10.06 ± 0.70 **	19.88 ± 0.37 **	0.71 ± 0.03 **	0.155 ± 0.005 **
MSA-GAN	9.37 ± 0.71 **	19.86 ± 0.58 **	0.75 ± 0.02 **	0.150 ± 0.008 **
SynDiff	7.57 ± 0.74 **	21.74 ± 0.60 **	0.80 ± 0.04 *	0.130 ± 0.006 **
MSFRPM	6.26 ± 0.86	23.53 ± 0.92	0.83 ± 0.03	0.100 ± 0.010

Note: Statistical significance was assessed via paired *t*-test comparing each method against MSFRPM on the test set. MAE values are scaled by 100 for better readability.

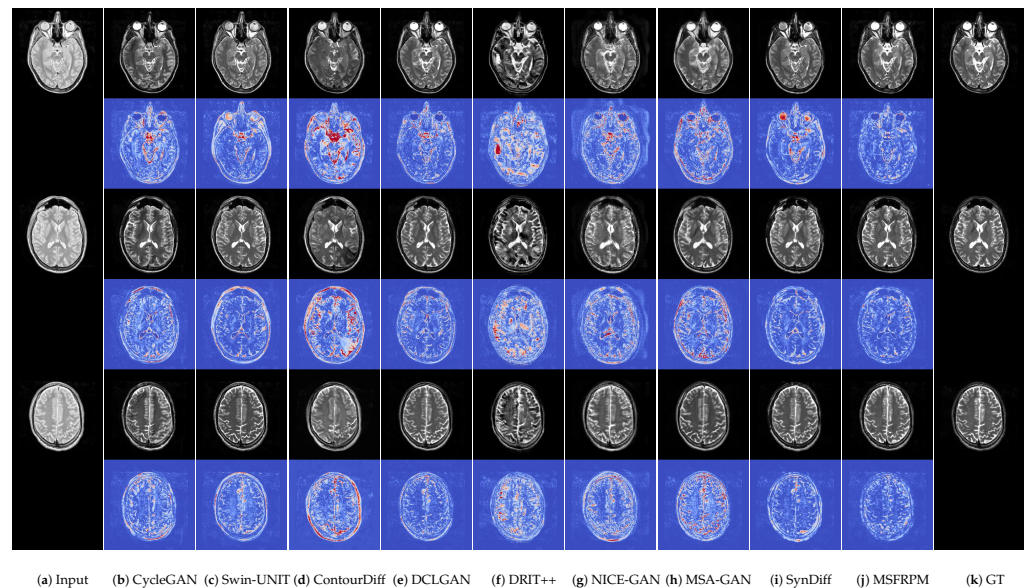


Figure 3. Visual results for brain PD-to-T2 MR image translation. The images in the even rows show heatmaps that reflect the difference between the synthetic images produced by each method and the ground truth. Warmer colors (red) indicate larger differences, while cooler colors (blue) represent smaller differences.

4.2. Model Analysis

In this section, we further investigate the effect of several key components of the proposed framework.

4.2.1. Effect of Multi-Scale Feature Regularization and Patch Mixup Strategy

We first investigate the effects of the proposed multi-scale feature regularization approach and the patch-based mixup strategy. To this end, we trained two variants of MSFRPM. One variant, labeled w/o MSFR (without Multi-Scale Feature Regularization), omits all short connections from the framework and applies feature regularization only to the lowest-level feature space. The other variant, labeled w/o PM (without Patch Mixup), excludes the patch-based mixup strategy.

Both variants were trained using the same procedure as MSFRPM. The quantitative results presented in Table 2 demonstrate the effectiveness of the proposed multi-scale feature regularization framework and the patch-based mixup strategy.

Table 2. Ablation study on key components of the proposed framework. Best results are marked in bold.

	MAE ($\times 100$) $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$
w/o MSFR	7.84 ± 0.87	21.17 ± 0.74	0.78 ± 0.04
w/o PM	6.66 ± 0.35	22.80 ± 0.37	0.82 ± 0.02
MSFRPM	6.26 ± 0.86	23.53 ± 0.92	0.83 ± 0.03

To further assess the impact of the proposed multi-scale feature regularization on cross-domain feature space alignment, we visually compare the features embedded by our MSFRPM model with those from its variant w/o MSFR. Specifically, we selected a pair of head PD and T2 MR images at random from the test dataset. Ideally, these paired images share the same anatomical structure, which should lead to consistent feature embeddings. We visualize the embedded features of T2 and PD images using the learned encoder, employing t-SNE [30] for dimensionality reduction. The results are illustrated in

Figure 4, where red and blue points represent features extracted from T2 and PD images, respectively. It is evident that the features produced by MSFRPM exhibit superior cross-domain alignment compared to those generated by the variant w/o MSFR. This further validates the effectiveness of our proposed multi-scale feature regularization approach.

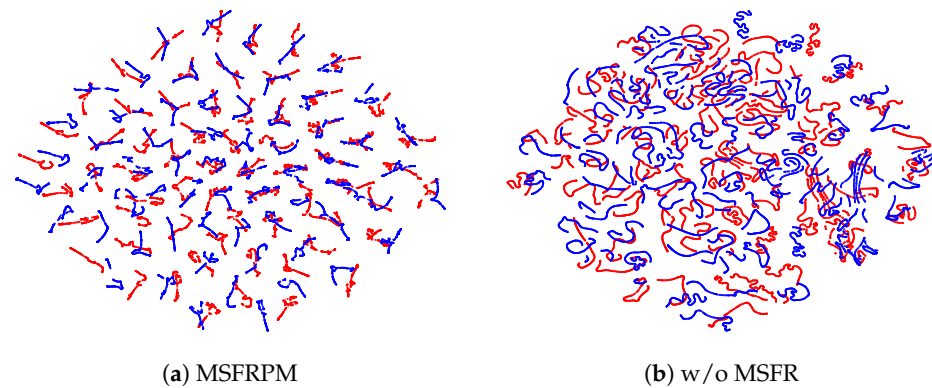


Figure 4. Visualization of features embedded from paired T2 (Red) and PD (Blue) images using t-SNE. (a) Feature visualization generated by MSFRPM. (b) Feature visualization generated by w/o MSFR.

4.2.2. Partition of High-Level and Low-Level Feature Groups

In our framework, the features generated by the intermediate layers of the encoder are categorized into the high-level group and low-level group. Therefore, a natural question is: how to partition the high-level group and low-level group? Let $\#\mathcal{H}$ denote the number of high-level scales (for example, in Figure 1, $\#\mathcal{H} = 3$). We conducted experiments to validate the impact of different $\#\mathcal{H}$ values on model performance. Figure 5 illustrates the performance with various partitions of high-level and low-level groups. We observed that $\#\mathcal{H} = 3$ yields the best results. Therefore, in this work, we categorize the last three encoding blocks as a high-level group and the remaining blocks as a low-level group, as illustrated in Figure 1.

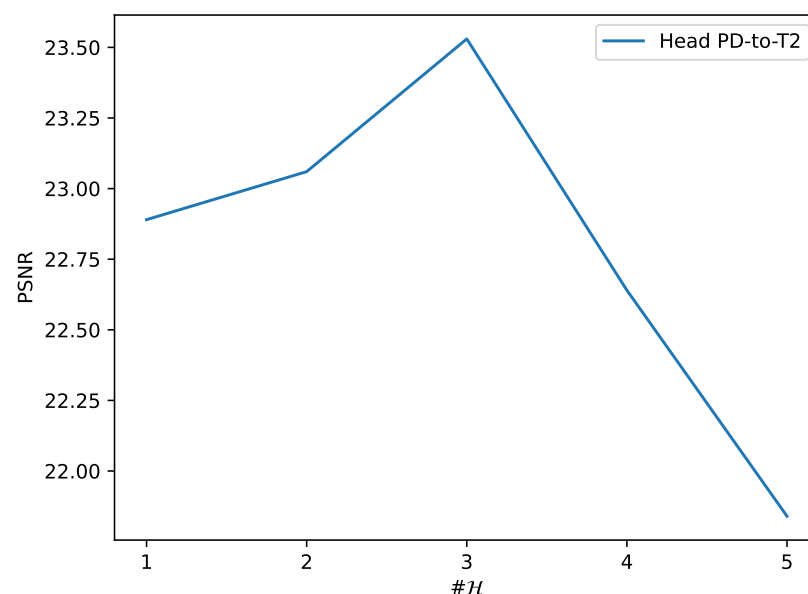


Figure 5. Performance with different partitions of high-level and low-level groups.

4.2.3. Analysis of Patch Mixup Strategy Hyper-Parameters

In the proposed patch-based mixup strategy, the mask used to blend images is controlled by two hyper-parameters: γ and η . Here, γ controls the expected proportion of spatial regions that are replaced after block expansion, while η controls the edge length of each square block. The block locations are not manually specified; instead, they are randomly sampled by Bernoulli trials over all spatial locations and then expanded into contiguous square regions. To assess the impact of different γ and η values on model performance, we trained the MSFRPM with various values of γ and η . The performance for different γ and η values is shown in Figures 6 and 7, respectively. The results indicate that, within the tested range, increasing γ generally improves model performance. For the hyper-parameter η , MSFRPM achieves the best performance at $\eta = 31$. Larger γ values lead to stronger masking and, after block expansion, may remove too much spatial content from the source image, which can in turn compromise structural fidelity. Considering this trade-off between augmentation strength and structural preservation, we set $\gamma = 0.5$ and $\eta = 31$.

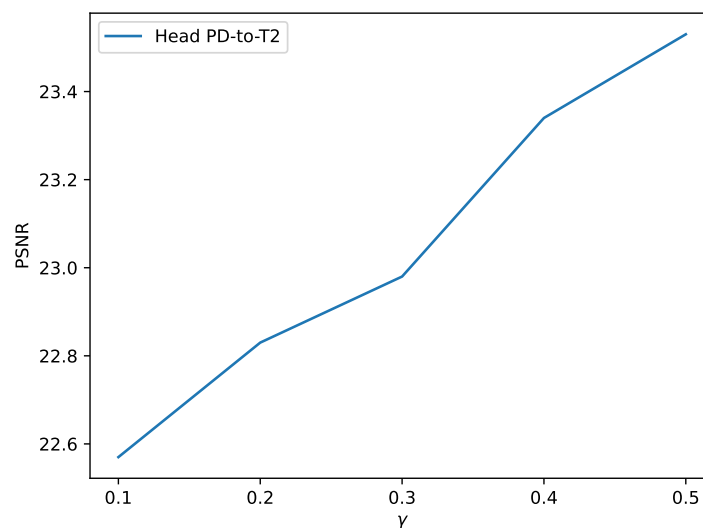


Figure 6. Performance with different γ .

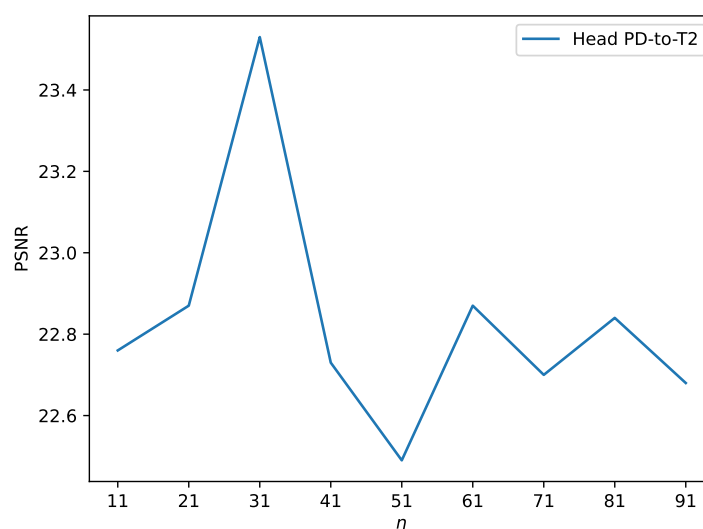


Figure 7. Performance with different η .

5. Discussion

In this study, we presented MSFRPM for unsupervised head PD-to-T2 MR image translation. Experimental results on the IXI dataset showed that our method performed favorably against eight comparison approaches across the reported metrics (MAE, PSNR, SSIM, and GMSD), with statistically significant differences ($p < 0.01$). The multi-scale feature regularization and patch mixup strategies effectively preserved low-level anatomical details while ensuring robust domain alignment, as validated by both quantitative metrics and t-SNE visualizations.

Because the IXI dataset contains healthy subjects only, the present experiments do not evaluate lesion preservation, pseudo-lesion risk, or downstream diagnostic performance. The current findings therefore support improved image fidelity and structural consistency on this dataset, while validation on pathological data remains future work.

In addition, the current implementation focuses on 2D slice-wise translation. While this approach is computationally efficient, it does not fully exploit 3D spatial contextual information or enforce inter-slice consistency. Moreover, the validation remains limited to a single public dataset of healthy subjects and does not include external cohorts, cross-site testing, or robustness analysis across scanners and acquisition conditions. Future work will therefore focus on 3D volumetric translation, multi-center validation, and downstream assessments on pathological data.

6. Conclusions

This study proposed MSFRPM, a novel unsupervised framework for head PD-to-T2 MR image translation. By introducing multi-scale feature regularization and a patch-based mixup strategy, the method achieved favorable results compared with the evaluated approaches, with improvements in MAE, PSNR, SSIM, and GMSD on the IXI dataset.

The proposed approach provides a useful framework for unsupervised PD-to-T2 image translation under the current experimental setting. Further validation on pathological and multi-center datasets is needed to assess its broader applicability.

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

Funding: This research was funded by National Natural Science Foundation of China (62276105) and Natural Science Foundation of Fujian Province (2023J01136).

Institutional Review Board Statement: The study utilized the publicly available IXI dataset, which was collected under the UK Engineering and Physical Sciences Research Council (EPSRC) funded project “Information extraction from Images” (EPSRC GR/S21533/02). As a major publicly funded research initiative, its data collection procedures adhered to the required ethical standards. For secondary analysis of such publicly released, anonymized data, no additional ethical approval is required.

Informed Consent Statement: Not applicable. The study utilized the publicly available IXI dataset with anonymized subject data.

Data Availability Statement: The data presented in this study are openly available in Zenodo at <https://doi.org/10.5281/zenodo.19249774>.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- De Guevara, M.L.; Echevarria, J.; Li, Y.; Hold-Geoffroy, Y.; Smith, C.; Ito, D. Cross-modal Latent Space Alignment for Image to Avatar Translation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 1–6 October 2023; pp. 520–529.
- Liu, M.Y.; Breuel, T.; Kautz, J. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc. (NeurIPS): San Diego, CA, USA, 2017; Volume 30.
- Huang, X.; Liu, M.Y.; Belongie, S.; Kautz, J. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 172–189.
- Lee, H.Y.; Tseng, H.Y.; Mao, Q.; Huang, J.B.; Lu, Y.D.; Singh, M.; Yang, M.H. Dri++: Diverse image-to-image translation via disentangled representations. *Int. J. Comput. Vis.* **2020**, *128*, 2402–2417. [\[CrossRef\]](#)
- Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
- Gu, X.; Zhang, Y.; Zeng, W.; Zhong, S.; Wang, H.; Liang, D.; Li, Z.; Hu, Z. Cross-modality image translation: CT image synthesis of MR brain images using multi generative network with perceptual supervision. *Comput. Methods Programs Biomed.* **2023**, *237*, 107571. [\[CrossRef\]](#) [\[PubMed\]](#)
- Hiasa, Y.; Otake, Y.; Takao, M.; Matsuoka, T.; Takashima, K.; Carass, A.; Prince, J.L.; Sugano, N.; Sato, Y. Cross-modality image synthesis from unpaired data using CycleGAN: Effects of gradient consistency loss and training data size. In *Simulation and Synthesis in Medical Imaging: Third International Workshop*; Springer: Cham, Switzerland, 2018; pp. 31–41.
- Armanious, K.; Jiang, C.; Abdulatif, S.; Küstner, T.; Gatidis, S.; Yang, B. Unsupervised medical image translation using cycle-MedGAN. In *2019 27th European Signal Processing Conference (EUSIPCO)*; IEEE: Piscataway, NJ, USA, 2019; pp. 1–5.
- Roy, P.K.; Santhanam, Misra, B.P.; Sen, A.; Palanisamy, T.; Gautam, S.; Krishna Murthy, S.; Arya, M. CycleGAN Implementation on Cross-Modality Transfer Between Magnetic Resonance Image (MRI) and Computed Tomography (CT) Images. In *Proceedings of the International Conference on Data Management, Analytics & Innovation*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 445–455.
- Han, J.; Shoeiby, M.; Petersson, L.; Armin, M.A. Dual contrastive learning for unsupervised image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 746–755.
- Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
- Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
- Guo, H.; Mao, Y.; Zhang, R. Augmenting data with mixup for sentence classification: An empirical study. *arXiv* **2019**, arXiv:1905.08941. [\[CrossRef\]](#)
- Verma, V.; Lamb, A.; Beckham, C.; Najafi, A.; Mitliagkas, I.; Lopez-Paz, D.; Bengio, Y. Manifold mixup: Better representations by interpolating hidden states. In *Proceedings of the International Conference on Machine Learning*; PMLR: Cambridge, MA, USA, 2019; pp. 6438–6447.
- Faramarzi, M.; Amini, M.; Badrinaaraayanan, A.; Verma, V.; Chandar, S. Patchup: A feature-space block-level regularization technique for convolutional neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*; Association for the Advancement of Artificial Intelligence (AAAI): Washington, DC, USA, 2022; Volume 36, pp. 589–597.
- Uddin, A.S.; Monira, M.S.; Shin, W.; Chung, T.C.; Bae, S.H. SaliencyMix: A Saliency Guided Data Augmentation Strategy for Better Regularization. In Proceedings of the 9th International Conference on Learning Representations, ICLR 2021, Virtual, 3–7 May 2021.
- Choi, H.K.; Choi, J.; Kim, H.J. Tokenmixup: Efficient attention-guided token-level data augmentation for transformers. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 14224–14235.
- Modat, M.; Cash, D.M.; Daga, P.; Winston, G.P.; Duncan, J.S.; Ourselin, S. Global image registration using a symmetric block-matching approach. *J. Med. Imaging* **2014**, *1*, 024003. [\[CrossRef\]](#) [\[PubMed\]](#)
- Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
- Karnewar, A.; Wang, O. Msg-gan: Multi-scale gradients for generative adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 7799–7808.
- He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *Proceedings of the European Conference on Computer Vision*; Springer: Cham, Switzerland, 2016; pp. 630–645.
- Wu, Y.; He, K. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*; Springer: Cham, Switzerland, 2018; pp. 3–19.

23. Ramachandran, P.; Zoph, B.; Le, Q.V. Searching for activation functions. *arXiv* **2017**, arXiv:1710.05941. [[CrossRef](#)]
24. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, Vancouver, BC, Canada, 30 April–3 May 2018.
25. Li, Y.; Li, Y.; Tang, W.; Zhu, Z.; Yang, J.; Liu, Y. Swin-UNIT: Transformer-based GAN for high-resolution unpaired image translation. In *Proceedings of the 31st ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 4657–4665.
26. Chen, Y.; Konz, N.; Gu, H.; Dong, H.; Chen, Y.; Li, L.; Lee, J.; Mazurowski, M.A. ContourDiff: Unpaired Image-to-Image Translation with Structural Consistency for Medical Imaging. *arXiv* **2024**, arXiv:2403.10786.
27. Chen, R.; Huang, W.; Huang, B.; Sun, F.; Fang, B. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 8168–8177.
28. Wang, F.; Zhang, Q.; Zhao, Q.; Wang, M.; Sun, F. Unsupervised image-to-image translation with multiscale attention generative adversarial network. *Appl. Intell.* **2024**, *54*, 6558–6578. [[CrossRef](#)]
29. Özbey, M.; Dalmaz, O.; Dar, S.U.; Bedel, H.A.; Öztürk, Ş.; Güngör, A.; Çukur, T. Unsupervised medical image translation with adversarial diffusion models. *IEEE Trans. Med. Imaging* **2023**, *42*, 3524–3539. [[CrossRef](#)] [[PubMed](#)]
30. Laurens, V.D.M.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.