Learning MRI Translation with Explicit Dynamic Texture and Structure Priors

Runyu Xiao, Zhangkai Ni, *Member, IEEE*, Junze Zhu, Hanli Wang, *Senior Member, IEEE*

***Abstract*— Magnetic Resonance Imaging (MRI) translation aims to convert images between modalities, enabling multi-modal data acquisition for downstream tasks such as disease diagnosis, tissue segmentation, and clinical decision-making. However, existing methods often struggle to accurately model anatomical details, limiting their ability to preserve fine structures and leading to suboptimal image quality. To address this limitation, we propose a novel MRI translation framework that integrates explicit texture and structure priors. By leveraging the strengths of both regression and generative models, our approach first employs a regression model to generate an initial translation. We then introduce a dynamically adaptive probability map to model complex textural and structural regions. This probability map, along with the initial translation, serves as a condition for a diffusion-based generative model, which refines anatomical details. Finally, we fuse the outputs of both models, guided by the probability map, to produce a high-quality translated image. Extensive experiments demonstrate the superiority of our method in preserving image fidelity and enhancing visual quality. The code will be publicly available.**

***Index Terms*— Medical image translation, texture and structure detection, magnetic resonance imaging, regression model, diffusion model**

I. Introduction

MULTI-MODAL medical imaging is essential for lesion assessment and diagnosis, as it provides complementary information from different modalities for a more comprehensive understanding of anatomical structures [1]. However, each imaging modality necessitates distinct technologies, protocols, and equipment, resulting in substantial costs and extended acquisition times [2]. To mitigate these challenges, medical image translation techniques have emerged as a promising solution [3]–[6]. Medical image translation aims to transform images between modalities using translation algorithms or neural networks [7]. This process typically involves learning-based methods to model the mapping between paired images of different modalities, enabling high-quality image translations that preserve diagnostic value [8], [9].

Runyu Xiao, Zhangkai Ni, and Junze Zhu are with the School of Computer Science and Technology and the Key Laboratory of Embedded System and Service Computing (Ministry of Education), Tongji University, Shanghai 200092, China (e-mail: 2332038@tongji.edu.cn; zkni@tongji.edu.cn; 2351114@tongji.edu.cn).

Hanli Wang is with the School of Electronics and Information Engineering, Tongji University, Shanghai 200092, China (e-mail: hanliwang@tongji.edu.cn).

Corresponding author: Zhangkai Ni (zkni@tongji.edu.cn)

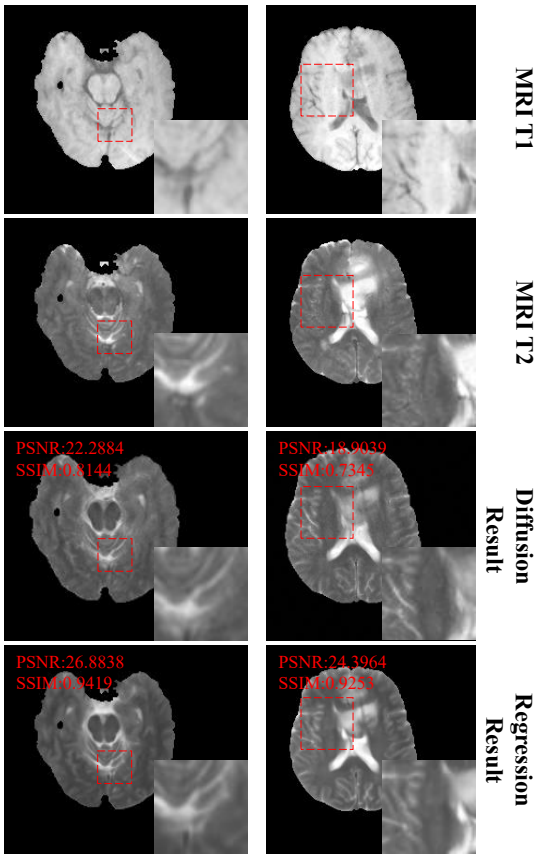

Fig. 1. From top to bottom: input T1-weighted MRI, ground-truth T2-weighted MRI, regression model outputs, and diffusion model outputs. The regression model attains higher PSNR and SSIM but fails to preserve fine anatomical details, whereas the diffusion model more faithfully restores missing textures and structures despite yielding lower quantitative scores.

Existing learning-based medical image translation methods can be broadly classified into two categories: Generative Adversarial Network (GAN)-based methods [10] and diffusion-based methods [11]. GAN-based approaches utilize a generator to translate images and a discriminator to differentiate between translated and ground-truth images. Although adversarial supervision is applied, these methods typically adopt an end-to-end encoder–decoder architecture, effectively functioning as regression models that directly map input images to their target counterparts. In contrast, diffusion models approximate the data distribution by gradually denoising samples from a noise-corrupted prior, enabling the step-wise generation of high-fidelity images.

TABLE I
QUANTITATIVE RESULTS ON THE BRATS DATASET. VALUES IN PARENTHESES INDICATE DIFFERENCES FROM THE REFERENCE T2-WEIGHTED IMAGES (SMALLER ABSOLUTE VALUE IS BETTER). FOR PSNR AND SSIM, HIGHER VALUES INDICATE BETTER QUALITY.

	MEAN	VAR	PSNR	SSIM
MRI T1	45.79	7222.60	-	-
MRI T2	28.22	2818.61	-	-
Regression Result	25.94 (-2.28)	2522.47 (-296.14)	27.23	0.9222
Diffusion Result	29.57 (+1.35)	2667.47 (-151.14)	26.35	0.9174

Both types of methods exhibit distinct strengths and limitations. As shown in Fig. 1, our empirical analysis reveals two key observations: 1) **Regression-based models tend to achieve higher objective evaluation metrics**, such as PSNR and SSIM. However, the resulting translated images often appear overly smooth and lack fine anatomical details. Moreover, these images tend to preserve structures from the input modality, leading to limited fidelity for the target modality. 2) **Diffusion-based models are more capable of synthesizing textures and structures that align closely with target images**, including missing anatomical details in the input. However, due to insufficient constraints during generation, these methods may hallucinate undesired anatomical content, adversely affecting objective quality metrics, leading to lower PSNR and SSIM scores. To validate these observations, we conduct a statistical analysis on the outputs of both model types using the BraTS dataset. As summarized in Tab. I, the results generated by the diffusion model show greater consistency with the target modality in terms of both intensity distribution (mean) and contrast (variance), indicating improved alignment in visual appearance [12]. In contrast, the regression-based model consistently yields higher quantitative scores, reinforcing the trade-off between visual realism and metric performance.

To overcome the respective limitations of regression and diffusion models, we propose to integrate their complementary strengths: generating missing anatomical structures while preserving faithful modality translation. This motivates a unified framework that produces high-quality translated images with both structural integrity and visual fidelity. The key challenges lie in: (1) identifying regions in the initially translated images that require refinement, and (2) effectively integrating the outputs of the regression and diffusion models. To address the first challenge, we introduce a texture- and structure-aware prior that localizes complex regions in the initial translation, which are then selectively refined using a diffusion-based model. For the second, we employ a dynamic, explicitly learned mask to adaptively fuse the outputs from both models via weighted summation, yielding the final high-quality translation. Our main contributions are summarized as follows:

- We conduct the first comprehensive qualitative and quantitative analysis of MRI T2-weighted images generated by different categories of medical image translation models, highlighting their respective strengths and limitations.
- Based on our observation and analysis, we propose a texture- and structure-aware evaluation metric to explicitly localize complex regions requiring refinement. A dynamic mask is then learned to adaptively fuse the outputs from the regression and diffusion models, resulting in high-fidelity, detail-preserving translations.
- Extensive experiments on benchmark datasets demonstrate that our method effectively synthesizes missing anatomical textures and structures while maintaining translation fidelity, achieving superior performance compared to existing state-of-the-art approaches.

II. RELATED WORKS

A. GAN-based Medical Image Translation Methods

The Generative Adversarial Network (GAN) was introduced by Goodfellow et al. [13], utilizing an adversarial learning process in which the generator and the discriminator engage in a Nash equilibrium game. This framework allowed the model to learn objectives that closely approximate the target data. Due to their superior visual performance, GANs have been widely applied in the field of medical image translation [14]. In 2015, Emily et al. [15] improved GANs by incorporating a Laplacian pyramid structure, enabling progressive high-resolution image generation. Phillip et al. [16] later introduced Pix2Pix, a conditional GAN using a PatchGAN discriminator for patch-level image translation. To address unpaired medical image translation, Zhu et al. [17] proposed CycleGAN, which leveraged cycle consistency for domain adaptation. Liu et al. [18] further extended the idea with UNIT, combining GANs and VAEs to learn a shared latent space for translation.

Since then, various GAN-based architectures have been proposed, focusing on lightweight design, stability, and high-quality image generation [5], [6], [19]–[21]. However, GANs still suffer from training instability and artifacts that degrade image quality. To address these limitations, recent studies suggest removing the discriminator and relying solely on a regression network, which improves training stability and output consistency. Accordingly, the first stage of our framework adopts an end-to-end encoder-decoder architecture for initial regression-based medical image translation.

B. Diffusion-based Medical Image Translation Methods

In recent years, diffusion models [22] have gained significant attention in the field of computer vision due to their outstanding generative capability and ability to model complex data distributions. Given that translation of medical image modality often involves misaligned features, diffusion models are particularly well suited for the tasks, as they can effectively generate missing textural and structural features [11]. Li et al. [23] used frequency filters to preserve structural details in CT translation. Muzaffer et al. [8] introduced Syndiff, an adversarial diffusion model that enhanced efficiency while ensuring high-fidelity translation. Kim et al. [24] combined SPADE with latent diffusion for modality-aware translation. Xing et al. [25] improved denoising by leveraging cross-conditional target modality distributions for more effective translation. Recently, Sanjeet et al. [26] introduced MMIT-DDPM, integrating text-based conditions and a structure-preserving module for joint denoising in medical image translation.

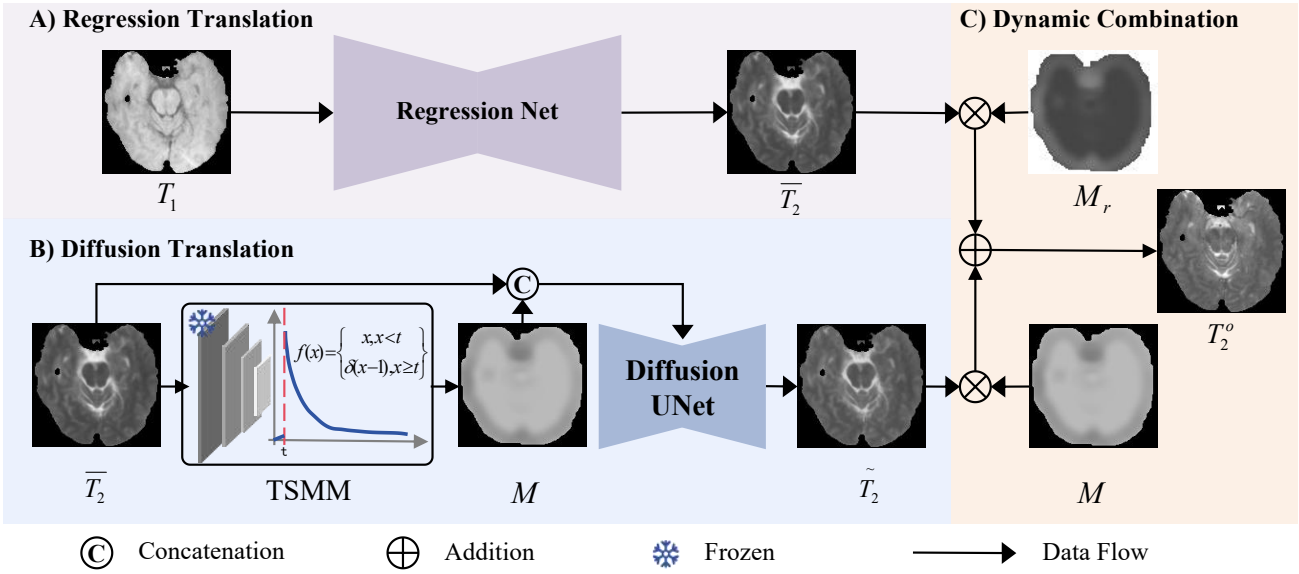

Fig. 2. The proposed framework comprises three key stages: **A) Regression-Based Translation.** The input MRI T1-weighted image $\boldsymbol{T_1}$ is first passed through a regression network to generate an initial translated image $\bar{\boldsymbol{T}}_{\boldsymbol{2}}$. **B) Diffusion-Based Refinement.** The initial result $\bar{\boldsymbol{T}}_{\boldsymbol{2}}$ is then processed by a Texture and Structure Measurement Module (TSMM), which produces a guidance mask $\boldsymbol{M}$. This mask, concatenated with $\bar{\boldsymbol{T}}_{\boldsymbol{2}}$, serves as the conditional input to a diffusion model to generate a refined image $\tilde{\boldsymbol{T}}_{\boldsymbol{2}}$. **C) Dynamic Fusion.** Finally, the mask $\boldsymbol{M}$ and its complement $\boldsymbol{M_r}$ are used to dynamically combine $\bar{\boldsymbol{T}}_{\boldsymbol{2}}$ and $\tilde{\boldsymbol{T}}_{\boldsymbol{2}}$, producing the final translated output $\boldsymbol{T}_{\boldsymbol{2}}^{\boldsymbol{o}}$.

While these methods have advanced medical image translation, they struggle to fully preserve complex anatomical details. Without sufficient constraints, diffusion models may introduce unwanted artifacts, potentially impacting downstream tasks like diagnosis and clinical decision-making. To overcome this limitation, we introduce a dynamic guidance mask in the second stage to steer the diffusion model toward critical textural and structural regions. Additionally, a dynamic fusion strategy is employed to adaptively combine outputs from both the regression and diffusion models, effectively suppressing artifacts and enhancing translation fidelity.

C. Textural and Structural Evaluation Metrics

Medical images often present complex textural and structural patterns, which introduce significant challenges for image restoration and translation models [27]. To address these challenges, effective evaluation metrics are needed to characterize such regions and guide model restoration. Huang et al. [28] leveraged first-order statistical features from auto-covariance matrices for liver lesion classification, while Ganeshan et al. [29] analyzed CT textures to assess tumor consistency. In MRI studies, Zacharaki et al. [30] demonstrated the effectiveness of texture and shape features for brain tumor classification, an approach echoed by Chen [31], Mayerhoefer [32], and others. Martin et al. [33] further emphasized texture in segmentation via a Mumford-Shah-based model. Inspired by perceptual quality assessment [34], [35], we propose leveraging multi-scale pre-trained feature maps to more effectively identify texture-rich regions and guide the translation model toward anatomically meaningful refinement.

III. Methodology

In this section, we first present the motivation and overall framework of our approach in Sec. III-A. Next, Sec. III-B provides a detailed description of the proposed Explicit Dynamic Textural and Structural Prior. Sec. III-C introduces the core components of our model and their interactions, along with how the explicit dynamic mask enables effective integration of the regression and diffusion models.

A. Overview

1) Motivation: : Given an MRI T1-weighted image T_1, our goal is to generate a corresponding T2-weighted image T_2^o that closely aligns with the target T_2. This translation task faces two main challenges:

- **Lack of explicit modeling for textural and structural regions causes distortions.** Most existing methods treat the entire image uniformly, either through regression or diffusion, without explicitly identifying complex anatomical regions. This often leads to missed details or unwanted artifacts, compromising the fidelity of the translated images.
- **Insufficient integration of complementary models causes sub-optimal results.** Cascading regression and diffusion models often under-utilize their strengths, while naive fusion can introduce inconsistencies and degrade reconstruction quality, which can cause noticeable inconsistencies, degrade visual quality, and potentially affect clinical interpretation.

To address these two issues, we propose a medical image translation method that leverages explicit modeling through textural and structural priors. First, a regression network produces an initial T2-weighted estimate, from which we derive a dynamic mask that highlights complex textural and structural regions. Next, this mask and the initial estimate jointly condition a diffusion model, directing its refinement to the most anatomically significant areas. Finally, we adaptively fuse

the regression and diffusion outputs to produce high-fidelity translations that preserve both visual quality and diagnostic relevance.

2) Framework: In this section, we provide an overview of the proposed framework, as illustrated in Fig. 2. Given an input MRI T1-weighted image $T_1 \in \mathbb{R}^{1\times H\times W}$, the goal is to generate the corresponding MRI T2-weighted image $T_2^o \in \mathbb{R}^{1\times H\times W}$, where H and W is the image height and width.

Specifically, T_1 is first processed by a regression network RegN$(\cdot)$, composed of convolutional layers and attention modules, to produce an initial translation $\bar{T_2}$:

$$\bar{T_2} = \text{RegN}(T_1). \tag{1}$$

Next, $\bar{T_2}$ is passed to the Texture and Structure Measurement Module TSMM$(\cdot)$ to extract an explicit structural-textural mask M:

$$M = \text{TSMM}(\bar{T_2}). \tag{2}$$

The mask M and initial translation $\bar{T_2}$ are concatenated channel-wise and fed into the diffusion model DIFF$(\cdot)$, producing a refined output $\tilde{T_2}$:

$$\tilde{T_2} = \text{DIFF}(\bar{T_2}||M), \tag{3}$$

where $||$ represents channel-wise concatenation.

Finally, the outputs of the regression and diffusion models are fused using a dynamic weighting strategy guided by M, yielding the final translated image:

$$T_2^o = M \times \tilde{T_2} + (1 - M) \times \bar{T_2}, \tag{4}$$

where $\times$ and $+$ denote pixel-wise multiplication and addition, respectively.

B. Explicit Dynamic Textural and Structural Prior

Regression-based networks offer reliable initial translations that mitigate the complexity of modeling source-to-target mappings. To enhance these preliminary results, a diffusion model is employed to iteratively refine the translations by incorporating dynamic textural and structural priors. To fully exploit the complementary strengths of both approaches, we draw inspiration from [34], [35] and explore the use of a pre-trained VGG network [36] to extract rich texture and structure cues from medical images. Following the practices in [37], [38], the input grayscale images are replicated across three channels to enable effective feature extraction using the pre-trained VGG network.

As shown in Fig. 3, the feature extraction capacity of the pre-trained VGG network varies with depth. As the depth increases, the features become more global, while the shallow features can better represent the textural and structural patterns of the image. Leveraging this property, we design a Texture and Structure Measurement Module (TSMM) to identify regions with pronounced textural and structural complexity in the initial translation from the regression network. By focusing on these key areas, TSMM directs the refinement process, enabling more accurate corrections and improving both the visual quality and consistency of the final output.

The TSMM primarily consists of two components: feature extraction and probability computation. For the input image

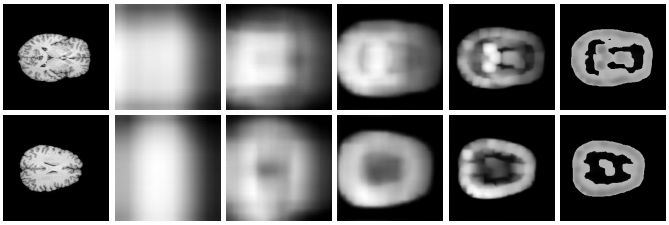

Fig. 3. From left to right: the input T1-weighted MRI image $\boldsymbol{T_1}$; feature probability maps from the pre-trained VGG network at A) layer 23, B) layer 16, C) layer 9, and D) layer 4; and the final dynamic mask map derived as described in Eq. 8. It can be observed that as the network layer deepens, the captured textural and structural information becomes increasingly global.

$\bar{T_2}$, multi-scale features are extracted from selected layers of a pre-trained VGG network, which can be defined as:

$$F_k = \text{VGG}_i(\bar{T_2}), i = [4, 8, 16, 23], k = 0, 1, 2, 3, \tag{5}$$

where i denotes the index of the VGG layer used for feature extraction, and k is the corresponding index of the extracted feature map used for subsequent computation.

For each feature map F_k, we compute the mean μ_k^c and variance σ_k^c across spatial dimensions for each channel c. These statistics are used to derive the dispersion index γ_k [39], which quantifies feature variability and is formulated as:

$$\begin{aligned}
\mu_k^c &= \frac{1}{H_k \times W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} F_k^{c,h,w}, \\
\sigma_k^c &= \frac{1}{H_k \times W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} \left(F_k^{c,h,w} - \mu_k^c\right)^2, \\
\gamma_k &= \frac{1}{C_k} \sum_{c=1}^{C_k} \frac{\sigma_k^c}{\mu_k^c + \epsilon},
\end{aligned} \tag{6}$$

where H_k and W_k denote the spatial dimensions of the k-th feature map, C_k is the number of channels, and ϵ is a small constant (set to $1e-12$) to ensure numerical stability.

Subsequently, the computed dispersion index γ_k is passed through a sigmoid function $\delta(\cdot)$, yielding a normalized probability score ps_k. This score is then element-wise multiplied with the upsampled probability map $\hat{p}_k$ from the subsequent deeper layer, which has been bilinearly interpolated to match the spatial resolution of the current feature map. This fusion operation enables the propagation and refinement of textural and structural probabilities across layers. The process is formally defined as:

$$\begin{aligned}
ps_k &= \delta(\gamma_k) = \frac{1}{1 + e^{-\gamma_k}}, \\
\hat{p_k} &= \Delta(p_{k+1}, (H_k, W_k)), \\
p_k &= ps_k \cdot \hat{p_k},
\end{aligned} \tag{7}$$

where $\Delta(\cdot, (\cdot, \cdot))$ denotes the bilinear interpolation operation, and H_k and W_k represent the height and width of the k-th feature map, respectively.

This method produces a set of probability maps at multiple scales, from which the final processed map p_0 is selected as the candidate mask. To enhance its discriminative capacity,

regions with high response intensities are further refined using an empirically defined threshold. This refinement process can be described as follows:

$$p(h,w) = \begin{cases} p_0(h,w) & p_0(h,w) < t, \\ \frac{e^{1-p_0(h,w)}}{1+e^{1-p_0(h,w)}} & p_0(h,w) \geq t, \end{cases} \tag{8}$$

where p_0 denotes the final probability map, (h, w) are spatial coordinates. The t is the threshold value, which is empirically set to 0.05.

The mask image generated by the above process serves as a probability map for evaluating the presence of textural and structural information in the initially translated image. As shown in the final column of Fig. 3, this method effectively captures and emphasizes the relevant textural and structural features of the image.

C. MRI Translation with Texture and Structure Priors

1) Regression Network: The regression network is designed to generate an initial high-fidelity translation, thereby reducing the complexity of learning in subsequent networks. Motivated by recent advances in medical image translation [40]–[42], we adopt a U-Net architecture as the backbone to exploit its strength in multi-scale feature extraction. To further enhance contextual representation, we integrate Transformer-based attention mechanisms into the encoder path, drawing from the benefits demonstrated in [14], [43], [44]. These attention modules facilitate the selective emphasis of salient features while suppressing irrelevant or redundant information, ultimately improving the translation quality. As illustrated in Fig. 4, the input T1-weighted image T_1 is first processed by a convolutional layer that transforms it into the feature domain:

$$x_0 = \text{Conv}(T_1). \tag{9}$$

The feature map x_0 is then sequentially passed through a stack of Transformer blocks and downsampling operations to extract hierarchical representations:

$$x_i = \text{DS}(\text{TB}_e(x_{i-1}), i \in [1, N], \tag{10}$$

where $\text{TB}_e(\cdot)$ denotes the Transformer block in the encoder, $\text{DS}(\cdot)$ represents the downsampling module, and N indicates the number of encoding layers.

Subsequently, the deepest encoder features x_N are passed through a latent module, which adopts a structure similar to the encoder, to extract enriched features $\bar{x}_N$. These features are then progressively upsampled and concatenated with the corresponding encoder features at each resolution level. The combined features are processed through a convolutional layer followed by decoding Transformer Blocks to obtain the decoded features, formulated as:

$$\bar{x}_{i-1} = \text{TB}_d(\text{Conv}(\text{US}(\bar{x}_i)||x_{i-1})), i \in [N, 1], \tag{11}$$

where $\text{US}(\cdot)$ denotes the upsampling operation, $\text{TB}_d(\cdot)$ is the Transformer Block in the decoder, and $||$ represents feature concatenation.

The final decoded features $\bar{x}_0$ are refined via a refinement module, followed by a convolutional layer to project them back into the image domain. A residual connection is then applied

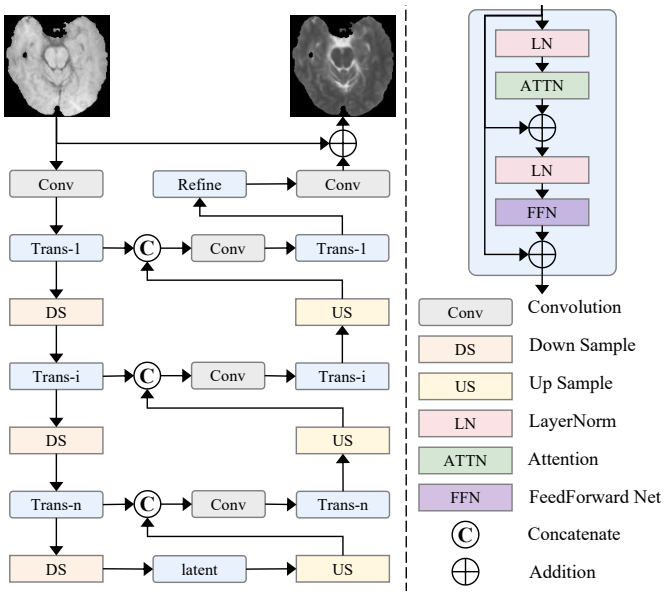

Fig. 4. Architecture of the regression network based on U-Net. The input T1-weighted image $\boldsymbol{T_1}$ is first mapped to the feature domain via convolution and then processed by a series of Transformer blocks, each comprising self-attention and a feed-forward network. Encoded features are fused with skip connections and decoded, followed by global refinement and residual connections to produce the final translation.

by adding the input image T_1 to the refined output, yielding the translated image $\bar{T}_2$:

$$\bar{T}_2 = T_1 + \text{Conv}(\text{RM}(\bar{x}_0)), \tag{12}$$

where $\text{RM}(\cdot)$ represents the Refinement Module, which shares the same structure as the Transformer Block.

The Transformer Block can be expressed as follows:

$$x_{out} = \text{FFN}(\text{LN}(x_{in} + \text{ATTN}(\text{LN}(x_{in}))) + x_{in}, \tag{13}$$

where $\text{LN}(\cdot)$, $\text{ATTN}(\cdot)$, and $\text{FFN}(\cdot)$ represent the Layer Normalization module, Attention module, and Feed Forward Network, respectively.

This design yields a high-fidelity preliminary translated image $\bar{T}_2$, which serves as the input for subsequent stages of the framework.

2) Diffusion Network: Diffusion models have gained significant attention for their remarkable generative capabilities. In this study, the diffusion model is utilized to refine the initial translated image $\bar{T}_2$, guided by dynamic texture and structure maps, aiming to produce high-quality translated outputs. As illustrated in Fig. 5, the diffusion process consists of two main steps: the Forward Process and the Reverse Process.

The forward process incrementally adds Gaussian noise to the input and is defined as:

$$\begin{aligned} x_t &= \sqrt{1-\beta_t} x_{t-1} + \sqrt{\beta_t}\epsilon, \\ q(x_t|x_{t-1}) &= \mathcal{N}(x_t; \sqrt{1-\beta_t} x_{t-1}, \beta_t I), \end{aligned} \tag{14}$$

where x_t represents the noisy feature at timestep t, which has the same dimensionality as the original data x_0. $\mathcal{N}(\cdot;\cdot,\cdot)$ denotes the Gaussian distribution, β_t is a predefined variance schedule at timestep t. ϵ and I denote the added noise and the identity covariance matrix, respectively.

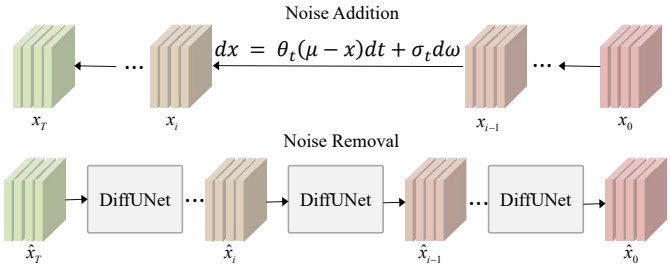

Fig. 5. The diffusion model consists of a forward process and a reverse process. **Forward Process**: According to the noise schedule, the image is progressively corrupted into Gaussian noise. **Reverse Process**: Through the network's learned denoising mechanism, the noise is gradually removed, and the image is progressively restored.

A neural network with the parameter θ is then used to perform the Reverse Process:

$$\begin{aligned} p_\theta(x_{0:T}) &:= p(x_T)\sum_{t=1}^{T} p_\theta(x_{t-1}|x_t), \\ p_\theta(x_{t-1}|x_t) &:= \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)), \end{aligned} \tag{15}$$

where $\mu_\theta(\cdot,\cdot)$ computes the mean while $\Sigma_\theta(\cdot,\cdot)$ calculates the variance, both using the neural network parameterized by θ.

The DiffUNet used in the Reverse Process shares a similar overall structure with the regression network, with the addition of a Time Embedding module that is tied to the diffusion time step. For further details, please refer to [22]. The diffusion model takes as input both the Mask Map M and the initial translated result $\bar{T}_2$. Through a gradual denoising process, the model generates the corrected diffusion output $\tilde{T}_2$, which exhibits improved textures and structures. $\tilde{T}_2$ is then used to generate the final output T_2^o.

IV. Experiments and Analysis

A. Experimental Setups

1) Datasets: To evaluate the performance of our model, we perform experiments on two publicly available datasets: the Brain Tumor Segmentation Challenge dataset (BraTS[1]) and the Information eXtraction from Images dataset (IXI[2]). The BraTS dataset consists of brain tumor MRI images with multiple modalities, including T1-weighted, T2-weighted, and FLAIR-weighted images. We adopt the data processing and partitioning procedure mentioned in [45], resulting in 5760 paired T1-weighted and T2-weighted images for training, and 768 paired images with the same modality for validation. All images have a resolution of 256 $\times$ 256 pixels. The IXI dataset, provided by the University College London (UCL) and collaborating institutions, includes a variety of medical imaging modalities. Following the data processing and partitioning method detailed in [46], we obtain 5268 paired T1-weighted and T2-weighted images for training and 660 pairs for validation. These images are padded with black borders to the resolution of 256 $\times$ 256 pixels to meet the requirements of the network.

[1]https://www.med.upenn.edu/sbia/brats2018/data.html
[2]https://brain-development.org/ixi-dataset/

2) Training Objective: The network is optimized using a combination of fidelity loss and SSIM loss, where fidelity loss is calculated using the L1 norm. Both fidelity loss and SSIM loss are evaluated between the initial translated result $\bar{T}_2$, the diffusion output $\tilde{T}_2$, the final translated result T_2^o, and the ground truth image T_2. The respective loss can be expressed as follows:

$$\begin{aligned} \mathcal{L}_{fid} &= |\bar{T}_2 - T_2| + |\tilde{T}_2 - T_2| + |T_2^o - T_2|, \\ \mathcal{L}_{ssim} &= (1 - \Gamma(\bar{T}_2, T_2)) + (1 - \Gamma(\tilde{T}_2, T_2)) \\ &\quad + (1 - \Gamma(T_2^o, T_2)), \end{aligned} \tag{16}$$

where $\Gamma(\cdot,\cdot)$ represents the SSIM calculation operation.

The total training loss of our model can be expressed as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{fid} + \alpha\mathcal{L}_{ssim}, \tag{17}$$

where α is the weighting factor controlling the relative importance of the fidelity loss and SSIM loss.

3) Evaluation Metrics: For quantitative comparison, classic metrics including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [48], and Learned Perceptual Image Patch Similarity (LPIPS) [49] are adopted to evaluate the quality of the translated images. PSNR primarily measures image fidelity, while SSIM emphasizes structural consistency. LPIPS, on the other hand, measures perceptual similarity based on deep neural networks, which aligns more closely with the human visual system (HVS). Notably, higher PSNR and SSIM values indicate better performance, whereas lower LPIPS scores correspond to smaller perceptual differences between images.

4) Training Details: **Settings of EDTS-Trans:** All input MRI T1-weighted images are preprocessed to a fixed resolution of 256 $\times$ 256 for network training. The paired MRI T1-weighted and T2-weighted images are augmented using identical random transformations, including horizontal and vertical flipping, 90-degree rotations, and their combinations. The initial learning rate is set to $2e-4$ and gradually decays to $1e-6$ following a cosine annealing schedule. Training is conducted for a total of 25,000 iterations to ensure sufficient convergence and fitting of the network. The model is implemented in PyTorch and trained on a single NVIDIA GeForce RTX 4090 GPU. The hyperparameter α in Equ. 17 is empirically set to 0.5. The number of feature map layers in Equ. 8 is set to 5, while the number of U-Net layers in Equ. 10 and Equ. 11 is set to 4. In addition, the input batch size is set to 1, consistent with most of the baseline settings.

Setting of Baselines: We conduct a comprehensive comparison between state-of-the-art GAN-based methods, diffusion-based methods, and our proposed model. Specifically, we include several representative image translation networks and recent diffusion-based approaches, such as UNIT [18], MUNIT [47], RegGAN [5], pGAN [16], Fast-DDPM [45], SynDiff [8], MMIT-DDPM [26], and ALDM [24]. For UNIT, MUNIT, and RegGAN, we adopt the official RegGAN implementation, modifying only the data loading path while keeping all other hyperparameters unchanged. For pGAN, we adjust the number of training epochs to ensure sufficient convergence on our dataset. For Fast-DDPM, ALDM, and SynDiff, we

TABLE II
RESULTS OF OUR PROPOSED EDTS-TRANS COMPARED WITH SOTA METHODS IN PSNR, SSIM, AND LPIPS. FOR EASE OF IDENTIFICATION, THE BEST- AND SECOND-BEST-PERFORMING ARE HIGHLIGHTED IN BOLD AND UNDERLINED, RESPECTIVELY.

Methods	Params (M)	FLOPS (G)	BraTS			IXI		
			PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
pGAN [16]	14.14	56.5	25.64	0.8556	0.0648	28.10	0.8505	<u>0.1033</u>
UNIT [18]	30.53	162.7	25.78	0.8577	0.0629	28.89	0.8597	0.1357
MUNIT [47]	38.29	154.5	25.13	0.8536	0.0683	28.00	0.8465	0.1454
Fast-DDPM [45]	48.46	132.6	26.30	0.9126	0.0634	29.17	0.9232	0.1344
RegGAN [5]	14.13	59.21	26.21	0.8612	0.0630	28.91	0.8564	0.1411
Syndiff [8]	55.32	160.8	<u>27.10</u>	<u>0.9235</u>	<u>0.0602</u>	<u>30.32</u>	<u>0.9328</u>	0.1213
MMIT-DDPM [26]	30.32	108.2	26.92	0.9228	0.0622	29.90	0.9104	0.1257
ALDM [24]	32.46	110.3	25.82	0.9040	0.0632	28.93	0.8762	0.1423
Ours	33.15	102.4	**27.60**	**0.9381**	**0.0578**	**30.65**	**0.9408**	**0.0977**

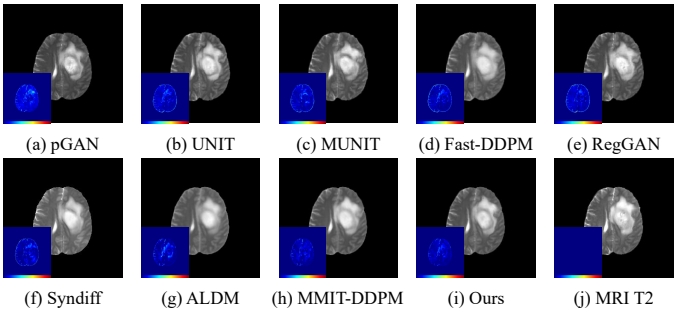

Fig. 6. The visual comparisons of the translation results of different methods for MRI images from the BraTS dataset. The error map at the left-bottom corner shows the residual between the results and the ground truth (GT), where brighter colors indicate a larger difference.

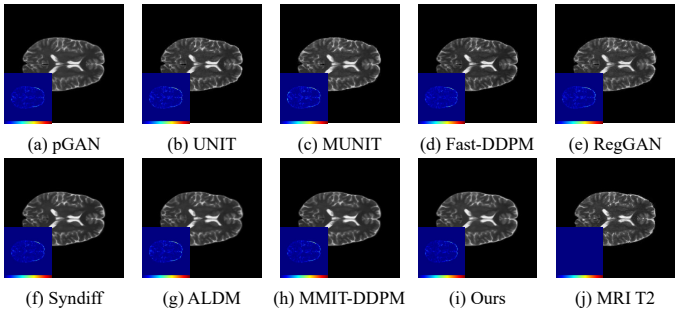

Fig. 7. The visual comparisons of the translation results of different methods for MRI images from the IXI dataset. The error map at the left-bottom corner shows the residual between the results and the ground truth (GT), where brighter colors indicate a larger difference.

follow the original experimental settings described in their respective papers, with only the batch size adjusted for fair comparison. For MMIT-DDPM, we modify the validation input size to 256 × 256 to maintain consistency with our experimental settings.

B. Comparison With State-of-the-Art Methods

1) Quantitative Comparison: The results presented in Tab. II clearly demonstrate the superior performance of our proposed method across a range of evaluation metrics, outperforming existing approaches. Specifically, on the BraTS dataset, our method achieves improvements of 0.50 in PSNR and 1.58% in SSIM over the state-of-the-art method, Syndiff. In the IXI dataset, our method shows notable gains of 0.33 in PSNR and 0.86% in SSIM. In addition, the corresponding LPIPS scores also exhibit a noticeable enhancement, further confirming the effectiveness of our approach. These results underscore the consistent high performance of our method across different datasets, reflecting its robustness and effectiveness in handling image translation tasks. Overall, our proposed method demonstrates its superiority by outperforming existing methods in fidelity and perceptual quality, demonstrating its ability to generate high-quality translated images.

2) Qualitative Comparison: Two representative sets of images from the BraTS and IXI datasets are selected to visually demonstrate the effectiveness of our model for MRI translation tasks. As shown in Fig. 6, our method generates visually compelling results on the BraTS dataset, with translated images that closely match the MRI T2-weighted ground truth and exhibit minimal artifacts. While some baseline methods also perform reasonably well, they often suffer from color bias or fail to simultaneously preserve texture, anatomical structure, and fidelity. Similarly, Fig. 7 illustrates that our approach yields consistent improvements on the IXI dataset. In addition, Fig. 8 presents t-SNE [50] visualizations for dimensionality reduction, comparing translated and ground truth T2-weighted images. The embeddings of our method align more closely with the ground truth compared to competing methods, further validating its effectiveness.

In summary, our method successfully integrates the advantages of regression-based and diffusion-based approaches to generate high-quality T2-weighted MRI images. The introduction of an explicit dynamic mask enables the model to maintain high fidelity while accurately capturing fine-grained textural and structural details, particularly in regions with complex anatomy. These results collectively demonstrate the robustness and effectiveness of our approach in challenging MRI image translation tasks.

C. Comparison on Segmentation Task

This subsection performs experiments on the translated results on the segmentation task to evaluate its effectiveness. Two widely used segmentation models, SAM2 [51] and MedSAM2 [52], are selected to validate our method. Zero-shot image segmentation is performed on MRI T1-weighted, MRI T2-weighted, and the translated results of our model. The

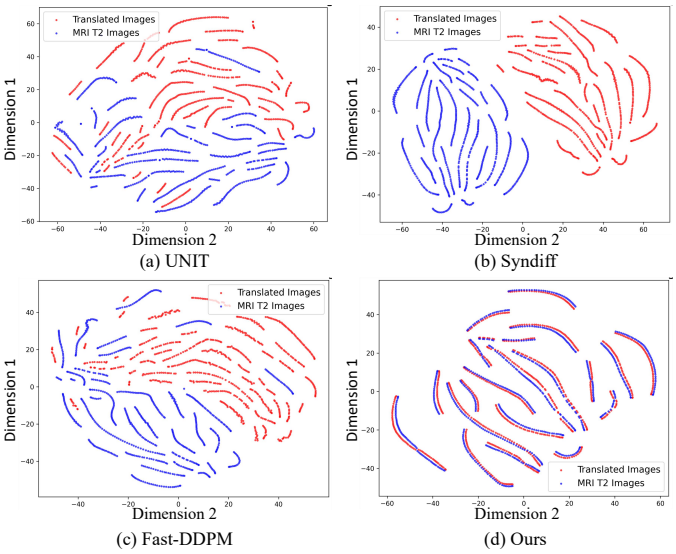

Fig. 8. The t-SNE results of different methods. The closer the blue and red dots are, the better the results are.

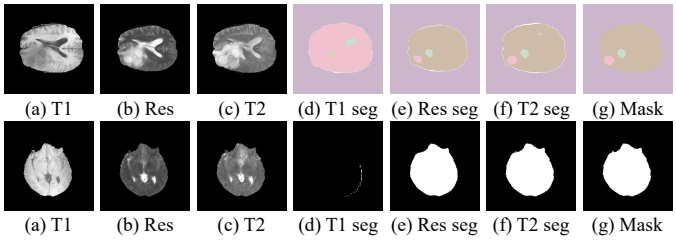

Fig. 9. Segmentation Results. The first row represents the first instance, where SAM2 is used for segmentation. It can be observed that MRI T2-weighted images achieve more accurate segmentation results. The second row corresponds to the second instance, where MedSAM2 is applied. Notably, images could not be segmented in MRI T1-weighted can be successfully segmented in MRI T2-weighted, demonstrating the advantages of MRI T2-weighted images for accurate segmentation.

mean Intersection over Union (mIoU) [53] and the dice coefficient [54] for T1-weighted images, T2-weighted images, and translated results are calculated with the ground truth masks. The detailed results are presented in the Tab. III. From the table, it can be observed that compared to MRI T1-weighted images, the translated results are closer to MRI T2-weighted images, achieving improvements in both evaluation metrics. Furthermore, as shown in the qualitative results in Fig. 9, MRI T2-weighted images enable more detailed segmentation compared to MRI T1-weighted ones, which can better assist physicians in diagnosis. Since the translated results closely resemble real MRI T2-weighted images, they can be effectively applied to downstream segmentation tasks.

D. Ablation Study

This subsection conducts extensive ablation studies to evaluate the effectiveness of various components proposed in our method, including the mask fusion method, dynamic mask weighting, and ablation experiments on the weight of loss function α.

1) Effectiveness of the proposed mask M*:* The first row in Tab. IV corresponds to the baseline approach in which

TABLE III
SEGMENTATION PERFORMANCE. HIGHER VALUES OF MIOU AND DICE COEFFICIENTS INDICATE BETTER SEGMENTATION RESULTS.

	Metric	SAM2 [51]	MedSAM2 [52]
T1	mIoU ↑	0.6051	0.5863
	DICE ↑	0.6490	0.5233
T2	mIoU ↑	0.6249	0.6728
	DICE ↑	0.6653	0.5905
Ours	mIoU ↑	0.6124	0.6237
	DICE ↑	0.6533	0.5468

TABLE IV
ABLATION STUDY OF THE MODULE COMPONENTS ON BRATS DATASET. W/O $\boldsymbol{M}$ REPRESENTS THE NETWORK WITHOUT EXPLICIT DYNAMIC MASK $\boldsymbol{M}$. TH-0.4 REPRESENTS THAT THE DYNAMIC MASK IS REPLACED BY THE MASK WITH A THRESHOLD EQUAL TO 0.4. W/O $\boldsymbol{L_{ssim}}$ REPRESENTS REMOVING SSIM LOSS.

Methods	PSNR↑	SSIM↑	LPIPS↓
w/o M	27.23	0.9222	0.0602
Th-0.4	27.39	0.9376	0.0587
w/o L_{ssim}	27.57	0.9371	0.0678
Ours	27.60	0.9381	0.0578

the regression network is applied first, and its output is directly fed into the diffusion model without incorporating the explicit dynamic masking strategy proposed in this work. The results show that excluding the masking strategy leads to a noticeable drop in performance. This degradation is likely due to the diffusion model's lack of guidance, which prevents it from performing targeted refinements, resulting in unnecessary unwanted artifacts in regions that do not require translation.

2) Effectiveness of dynamic fusion: The second row in Tab. IV represents the use of a threshold-based mask instead of the dynamic mask proposed in this paper. In this case, the output of the diffusion model is used for regions with values greater than 0.4, while the output of the regression model is applied to regions with values less than 0.4. The performance drop is attributed to the lack of dynamic weighting, which prevents the effective fusion of two results, leading to discontinuities. Additionally, the manually set threshold lacks adaptability to different examples. A comparison with the fourth row shows that the dynamic masking approach effectively bridges these discrepancies, resulting improvements.

3) Effectiveness of the loss function: The third row in Tab. IV represents the result obtained by setting the weight α of L_{ssim} to 0. As seen in the comparison with the fourth row, there is almost no significant difference in the evaluation metrics between the two. However, the presence of L_{ssim} helps to better constrain the structural information in the final image, thus improving the visual quality of the output. Therefore, in the final setting of our method, we retain a certain proportion of L_{ssim} weight.

E. Discussion

This subsection discusses the clinical relevance, innovation, and limitations of the proposed method, as evidenced by the

experimental results. MRI translation facilitates cross-modality image synthesis, enabling clinicians to infer missing contrasts without the need for additional scans. This capability reduces patient burden, shortens scan time, and reduces healthcare costs. Furthermore, it supports multi-contrast analysis for diagnosis and treatment planning when only single-modality data are available. The proposed method surpasses state-of-the-art approaches in both subjective visual quality, objective evaluation, and demonstrates strong performance in downstream segmentation tasks, validating its effectiveness. Distinct from existing approaches, our method explicitly models complex anatomical regions and innovatively integrates the strengths of regression and diffusion models via a dynamic texture-structure mask, enabling the generation of high-quality translated images. However, in certain instances, the translated outputs still present opportunities for refinement. Moreover, the current model exhibits limitations in inference efficiency when applied to high-resolution inputs. Addressing these challenges will constitute an important direction for our future research.

V. Conclusions

In this paper, we introduce a novel approach for MRI image translation, *i.e.* EDTS-Trans, which leverages explicit dynamic mask-guided modeling to improve the translation process. By incorporating a pre-trained model to extract textural and structural information from the initial translated images, the strengths of regression networks and diffusion models are effectively combined. This combination allows for the generation of translated MRI images that not only maintain high fidelity to the original images but also exhibit complex textural and structural details, which are crucial for accurate medical analysis. Extensive experiments conducted on two widely used public datasets demonstrate that EDTS-Trans consistently outperforms existing state-of-the-art methods in both qualitative and quantitative metrics. Moreover, the ablation studies highlight the effectiveness and robustness of architecture design and the dynamic mask-guided approach in improving the image translation task. In future work, we aim to further explore MRI image translation tasks, with an emphasis on meeting the requirements of downstream applications and medical diagnostics to ensure practical applicability.

References

[1] J. E. Iglesias, E. Konukoglu, D. Zikic, B. Glocker, L. K. Van, and B. Fischl. Is synthesizing MRI contrast useful for inter-modality analysis? In *Medical Image Computing and Computer-Assisted Intervention*, pages 631–638. Springer, 2013.

[2] S. Kaji and S. Kida. Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging. *Radiological Physics and Technology*, 12(3):235–248, 2019.

[3] K. Armanious, C. M. Jiang, S. Abdulatif, T. Küstner, S. Gatidis, and B. Yang. Unsupervised medical image translation using cycle-MedGAN. In *European Signal Processing Conference*, pages 1–5. IEEE, 2019.

[4] K. Armanious, C. M. Jiang, M. Fischer, T. Küstner, T. Hepp, K. Nikolaou, S. Gatidis, and B. Yang. MedGAN: medical image translation using GANs. *Computerized Medical Imaging and Graphics*, 79:101684, 2020.

[5] L. K. Kong, C. Y. Lian, D. T. Huang, Y. L. Hu, Q. C. Zhou, et al. Breaking the dilemma of medical image-to-image translation. *Advances in Neural Information Processing Systems*, 34:1964–1978, 2021.

[6] S. A. Yan, C. Y. Wang, W. B. Chen, and J. Lyu. Swin Transformer-based GAN for multi-modal medical image translation. *Frontiers in Oncology*, 12:942511, 2022.

[7] P. Paavilainen, S. U. Akram, and J. Kannala. Bridging the gap between paired and unpaired medical image translation. In *MICCAI Workshop on Deep Generative Models*, pages 35–44. Springer, 2021.

[8] M. Özbey, O. Dalmaz, S. U. Dar, H. A. Bedel, Ş. Özturk, A. Güngör, and T. Çukur. Unsupervised medical image translation with adversarial diffusion models. *Transactions on Medical Imaging*, 42(12):3524–3539, 2023.

[9] Y. F. Chen, Y. L. Lin, X. D. Xu, J. Z. Ding, C. Z. Li, Y. M. Zeng, W. F. Xie, and J. L. Huang. Multi-domain medical image translation generation for lung image classification based on generative adversarial networks. *Computer Methods and Programs in Biomedicine*, 229:107200, 2023.

[10] X. Yi, E. Walia, and P. Babyn. Generative adversarial network in medical imaging: a review. *Medical Image Analysis*, 58:101552, 2019.

[11] A. Kazerouni, E. K. Aghdam, M. Heidari, R. Azad, M. Fayyaz, I. Hacihaliloglu, and D. Merhof. Diffusion models in medical imaging: a comprehensive survey. *Medical Image Analysis*, 88:102846, 2023.

[12] M. Caballo, W. B. Sanderink, L. Y. Han, Y. Gao, A. Athanasiou, and R. M. Mann. Four-dimensional machine learning radiomics for the pre-treatment assessment of breast cancer pathologic complete response to neoadjuvant chemotherapy in dynamic contrast-enhanced MRI. *Journal of Magnetic Resonance Imaging*, 57(1):97–110, 2023.

[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *International Conference on Neural Information Processing Systems*, pages 2672–2680, 2014.

[14] N. K. Singh and K. Raza. Medical image generation using generative adversarial networks: a review. *Health Informatics: A Computational Perspective in Healthcare*, pages 77–96, 2021.

[15] E. Denton, S. Chintala, A. Szlam, and R. Fergus. Deep generative image models using a Laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1486–1494, 2015.

[16] P. Isola, J. Y. Zhu, T. H. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017.

[17] J. Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017.

[18] M. Y. Liu, T. Breuel, and J. Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[19] J. X. Chen, J. Wei, and R. Li. TarGAN: target-aware generative adversarial networks for multi-modality medical image translation. In *Medical Image Computing and Computer Assisted Intervention*, pages 24–33. Springer, 2021.

[20] Y. H. Ma, J. Liu, Y. H. Liu, H. Z. Fu, Y. Hu, J. Cheng, H. Qi, Y. F. Wu, J. Zhang, and Y. T. Zhao. Structure and illumination constrained GAN for medical image enhancement. *IEEE Transactions on Medical Imaging*, 40(12):3955–3967, 2021.

[21] S. Z. Yao, J. H. Tan, Y. Chen, and Y. H. Gu. A weighted feature transfer GAN for medical image synthesis. *Machine Vision and Applications*, 32(1):22, 2021.

[22] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[23] Y. X. Li, H. C. Shao, X. Liang, L. Y. Chen, R. Q. Li, S. Jiang, J. Wang, and Y. Zhang. Zero-shot medical image translation via frequency-guided diffusion models. *IEEE Transactions on Medical Imaging*, 43(3):980–993, 2023.

[24] J. H. Kim and H. Park. Adaptive latent diffusion model for 3D medical image to image translation: multi-modal magnetic resonance imaging study. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7604–7613, 2024.

[25] Z. H. Xing, S. C. Yang, S. X. Chen, T. Ye, Y. J. Yang, J. Qin, and L. Zhu. Cross-conditioned diffusion model for medical image to image translation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 201–211. Springer, 2024.

[26] S. S. Patil, R. Rajak, M. Ramteke, and A. S. Rathore. MMIT-DDPM–multilateral medical image translation with class and structure

supervised diffusion-based model. *Computers in Biology and Medicine*, 185:109501, 2025.

[27] M. K. Ghalati, A. Nunes, H. Ferreira, P. Serranho, and R. Bernardes. Texture analysis and its applications in biomedical imaging: a survey. *IEEE Reviews in Biomedical Engineering*, 15:222–246, 2021.

[28] Y. Huang, J. Chen, and W. Shen. Diagnosis of hepatic tumors with texture analysis in nonenhanced computed tomography images. *Academic Radiology*, 13(6):713–720, 2006.

[29] B. Ganeshan, K. Skogen, I. Pressney, D. Coutroubis, and K. Miles. Tumour heterogeneity in oesophageal cancer assessed by CT texture analysis: preliminary evidence of an association with tumour metabolism, stage, and survival. *Clinical Radiology*, 67(2):157–164, 2012.

[30] E. I. Zacharaki, S. M. Wang, S. Chawla, Y. D. Soo, R. Wolf, E. R. Melhem, and C. Davatzikos. Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 62(6):1609–1618, 2009.

[31] W. J. Chen, M. L. Giger, H. Li, U. Bick, and G. M. Newstead. Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. *Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine*, 58(3):562–571, 2007.

[32] M. E. Mayerhoefer, P. Szomolanyi, D. Jirak, A. Materka, and S. Trattnig. Effects of mri acquisition parameter variations and protocol heterogeneity on the results of texture analysis and pattern discrimination: an application-oriented study. *Medical Physics*, 36(4):1236–1243, 2009.

[33] M. Kiechle, M. Storath, A. Weinmann, and M. Kleinsteuber. Model-based learning of local image features for unsupervised texture segmentation. *IEEE Transactions on Image Processing*, 27(4):1994–2007, 2018.

[34] K. Y. Ding, K. D. Ma, S. Q. Wang, and E. P. Simoncelli. Image quality assessment: unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2020.

[35] K. Y. Ding, Y. Liu, X. Y. Zou, S. Q. Wang, and K. D. Ma. Locally adaptive structure and texture similarity for image quality assessment. In *International Conference on Multi-Media*, pages 2483–2491, 2021.

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[37] Q. S. Yang, P. K. Yan, Y. B. Zhang, H. Y. Yu, Y. Y. Shi, X. Q. Mou, M. K. Kalra, Y. Zhang, L. Sun, and G. Wang. Low-dose CT image denoising using a generative adversarial network with wasserstein distance and perceptual loss. *IEEE Transactions on Medical Imaging*, 37(6):1348–1357, 2018.

[38] A. Longuefosse, B. D. Senneville, G. Dournes, I. Benlala, P. Desbarats, and F. Baldacci. On the use of perceptual loss for fine structure generation: illustration on lung MR to CT synthesis. In *2024 IEEE International Symposium on Biomedical Imaging*, pages 1–5. IEEE, 2024.

[39] D. R. Cox and P. A. Lewis. The statistical analysis of series of events. *Monographs on Applied Probability and Statistics*, 1966.

[40] X. X. Yin, L. Sun, Y. H. Fu, R. L. Lu, and Y. C. Zhang. Retracted: U-Net-based medical image segmentation. *Journal of Healthcare Engineering*, 2022(1):4189781, 2022.

[41] R. Azad, E. K. Aghdam, A. Rauland, Y. W. Jia, A. H. Avval, A. Bozorgpour, S. Karimijafarbigloo, J. P. Cohen, E. Adeli, and D. Merhof. Medical image segmentation review: the success of U-Net. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10076–10095, 2024.

[42] J. N. Chen, J. R. Mei, X. H. Li, Y. Y. Lu, Q. H. Yu, Q. Y. Wei, X. D. Luo, Y. T. Xie, E. Adeli, Y. Wang, et al. TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of Transformers. *Medical Image Analysis*, 97:103280, 2024.

[43] K. Han, A. Xiao, E. H. Wu, J. Y. Guo, C. J. Xu, and Y. H. Wang. Transformer in Transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021.

[44] A. Gillioz, J. Casas, E. Mugellini, and A. K. Omar. Overview of the Transformer-based models for NLP tasks. In *Conference on Computer Science and Information Systems*, pages 179–183. IEEE, 2020.

[45] H. X. Jiang, M. Imran, L. H. Ma, T. Zhang, Y. Y. Zhou, M. X. Liang, K. Gong, and W. Shao. Fast-DDPM: fast denoising diffusion probabilistic models for medical image-to-image generation. *arXiv preprint arXiv:2405.14802*, 2024.

[46] J. H. Cole, R. P. Poudel, D. Tsagkrasoulis, M. W. Caan, C. Steves, T. D. Spector, and G. Montana. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage*, 163:115–124, 2017.

[47] X. Huang, M. Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision*, pages 172–189, 2018.

[48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[49] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.

[50] L. Van and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11), 2008.

[51] N. Ravi, V. Gabeur, Y. T. Hu, R. H. Hu, C. Ryali, T. Y. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, et al. SAM2: segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[52] J. Y. Zhu, A. Hamdi, Y. L. Qi, Y. M. Jin, and J. D. Wu. Medical SAM2: segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

[53] M. A. Rahman and Y. Wang. Optimizing intersection-over-union in deep neural networks for image segmentation. In *International Symposium on Visual Computing*, pages 234–244. Springer, 2016.

[54] F. Milletari, N. Navab, and S. Ahmadi. V-net: fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision*, pages 565–571. IEEE, 2016.